**QWE Customer Churn Analysis**

***Based on Machine Learning and Deep Learning Methods***

Zhou Jijia 20307100030

1. **QWE Case Overview**
2. Background of QWE Case

Customer churn has been a headache for many of corporations, especially for emerging Internet corporations. QWE is no exception. The company runs its business by managing the digital business performance of mid-small enterprises and charging service fees with subscriptions. As a mid-early-stage company struggling to break through the churn spell, the QWE managers determined to replace its passive strategy with proactive ones. Currently, facing with churn customers, QWE offers discounted or free services to unsubscribing customers to persuade them to stay. This method has mainly two drawbacks. First, trying to keep churning customers afterwards has a low probability to succeed since they have made up their mind to leave. Second, offering free services or discounts would largely affect the income stream of QWE. Thus, the current strategy is to be abandoned.

1. Goals and Preliminary Analysis

Under this circumstance, QWE decides to take the initiative and alter to a proactive customer churn solution. To be specific, QWE designed a customer retention strategy with 2 steps. The first step is to identify the possible churning customers. After that, QWE could make well-directed improvements on these customers in advance to their churn. During the early stage of this strategy implementation, emphasis will be put on churn customer prediction. The main goal will be designing an optimal model that could help QWE solve the churn problem to the largest extent. Subsidiary goals include figuring out the most influencing factors leading customers to churn and thereby providing instructions on service improvements.

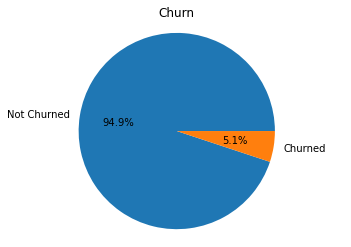
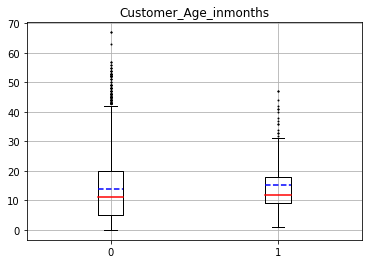
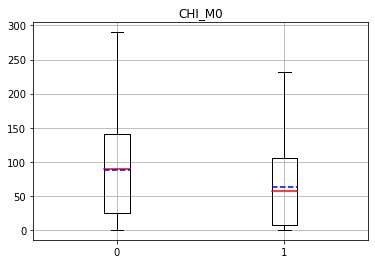
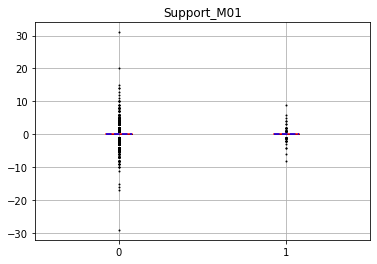
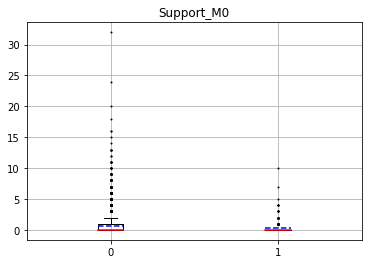
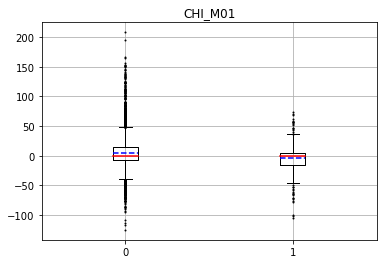
To simplify the problem and make our goal clearer, consider the following setting. Suppose QWE now has customers, and percent of them will churn if no or passive action is taken. Every churning customer would incur a loss of , and taking proactive actions to keep one customer would generate a cost of . QWE will spend money to save every customer predicted to churn and ignore the others. After several steps of mathematical derivation, the total cost can be determined by following function (Detailed derivation is in Appendix 1).

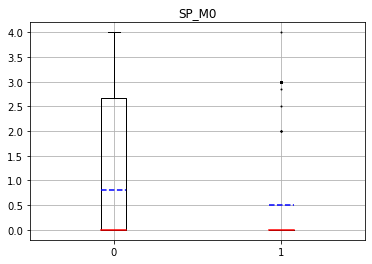
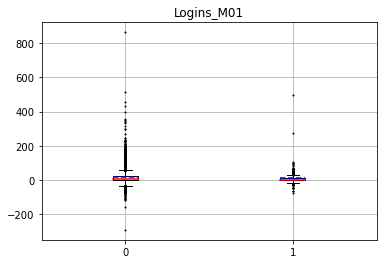
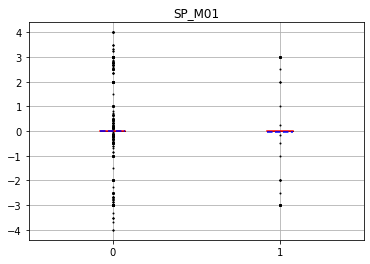
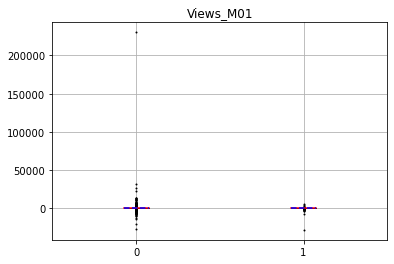
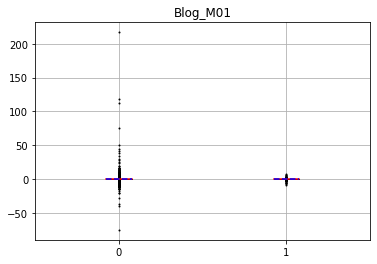
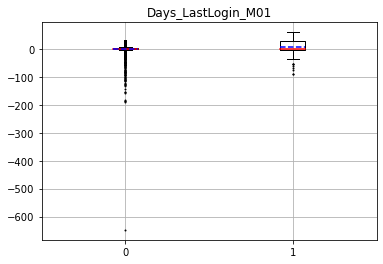
In this case, the cost of churn strategy is corresponding to both accuracy and recall, indicating that the predictor should optimize both accuracy and recall. Furthermore, the priority order of this two metrics is dependent on . For instance, if the loss of a churn customer is 21 times more than the customer-saving cost, and 10 percent of customers will churn. In this case, and , the total cost should be , implying that, in this case, recall is prior to overall accuracy if we want to minimize the total cost. The analysis above indicates that our goal is not merely finding a model with the best predicting performance, but also tuning it to a solution-oriented predictor with flexible accuracy and recall rate to satisfy the real need of QWE.

1. **QWE Dataset Analysis**
2. Variable Explanation and Preprocessing

In the dataset collected by QWE, there are 5077 samples in the train set, 4817 negatives and 260 positives respectively. Every sample has 12 variables, including one target variable and 11 predictor variables. They are Churn, Customer\_Age\_inmonths, *CHI\_M0, CHI\_M01, Support\_M0, Support\_M01, SP\_M0, SP\_M01, Logins\_M01, Blog\_M01, Views\_M01* and *DaysLastLogin\_M01.*Detailed explanations of these variables are listed in Appendix 2 along with their types.

The distributions of the variables are plotted as follows. The pie plot of *Churn* distribution shows that the dataset is a highly imbalanced dataset with rare samples of churned customers. The independent variable distributions of label 0 and 1 are plotted separately to observe the distinction better. Most boxplots demonstrate dispersed pattern of distribution, while *Customer\_Age\_inmonths* and *CHI\_M0* are concentrated. Some of the outliers, for example, the sample with *Views\_M01* larger than 200000, are considered erroneously sampled and therefore deleted.

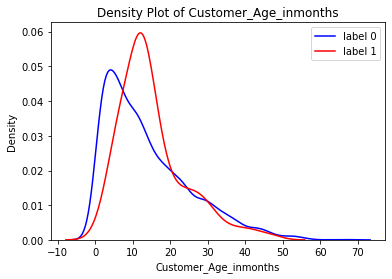
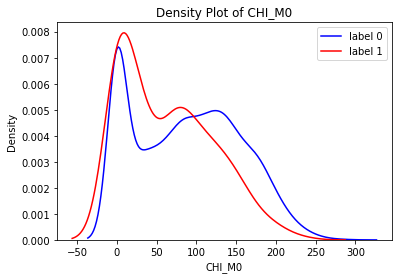
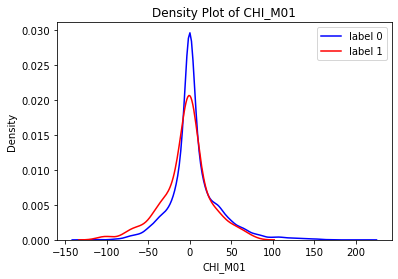
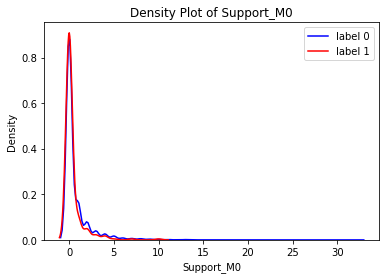
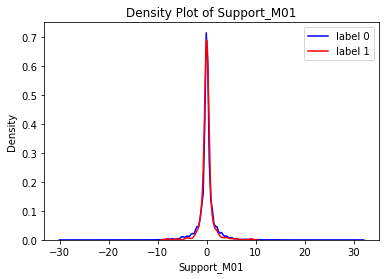
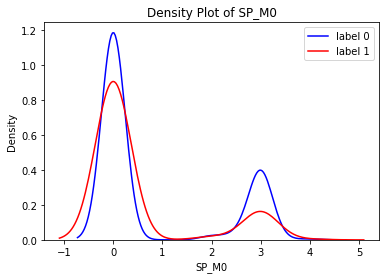
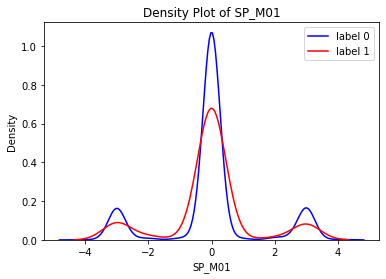
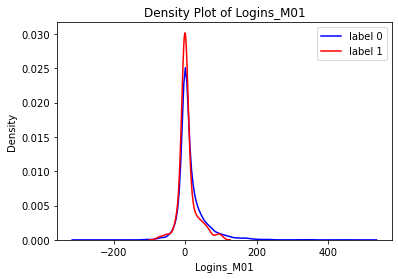
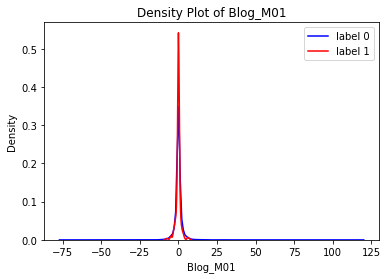
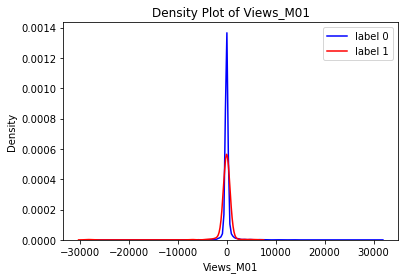
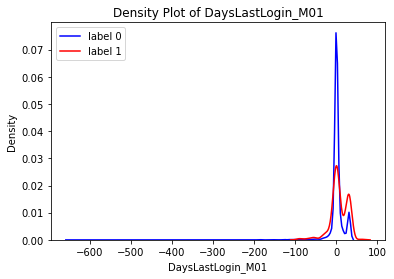
   

**Figure 1**. Variable distributions - pie plot and box plots

1. Explorative Data Analysis (EDA)

To illustrate the distribution of both negative and positive samples after cleaning, density plots of every variable are drawn. According to these plots, the negative and positive samples are approximately similar in distributions of *Support\_M0, Support\_M01, Logins\_M01* and *Blog\_M01*. On *CHI\_M01, SP\_M0, SP\_M01* and *Views\_M01*, the distribution patterns are alike but negative are more concentrated on certain values. While on *Customer\_Age\_inmonths* and *CHI\_M0*, the two distributions are distinct. Among the distinct variables we may extract the most effective features in following tasks.

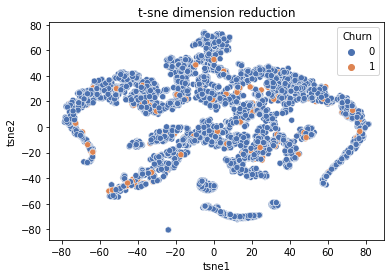
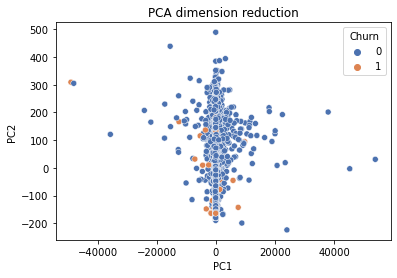
**Figure 2**. Predictor variable distributions after cleaning - density plot

Then, the correlation heat map clearly summarizes the correlations between the variables. First, the second column illustrates the correlation between the target variable with others. Warm color indicates positive correlation to the label while cold color indicates negative ones. Furthermore, *CHI\_M0, CHI\_M01, SP\_M0, Logins\_M01* and *DaysLastLogin\_M01* have relevantly stronger correlation with the target than others. Second, other columns imply the existence of multicollinearity. Taking *CHI\_M0* as an example, it is strongly positively correlated to *Customer\_Age\_inmonths* and *SP\_M0*. This might be because higher priority leads to satisfaction and higher satisfaction leads to longer time of subscription. These correlations between independent variables suggest that dimension reduction and feature selection might be good options for feature engineering.



**Figure 3**. Correlation heat map

Finally, to pre-evaluate the difficulty of distinguishing negative and positive samples, a classification attempt of dimension reduction is made. With Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), the train set dimension was reduced from 11 to 2. However, projected on 2-dimension space, the positive samples could not be clearly separated from negative ones. In either means, whether linear or non-linear, the two types of samples are still mixed together, implying that the classification task is complicated and require complex predictor models.

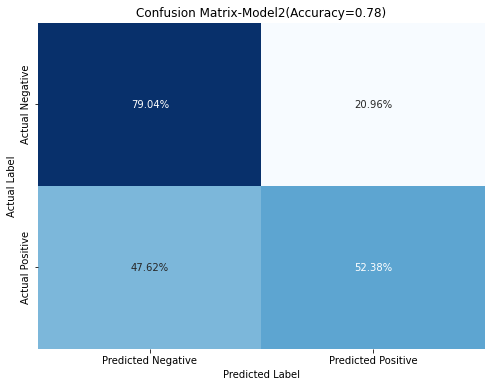
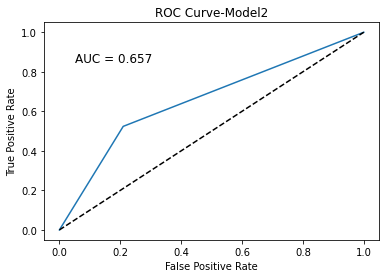


**Figure 4**. Dimension Reduction Visualization- PCA and t-SNE

1. **Predictor Models Designing and Interpretations**
2. Baseline and Explanatory Models: Logistic Regression

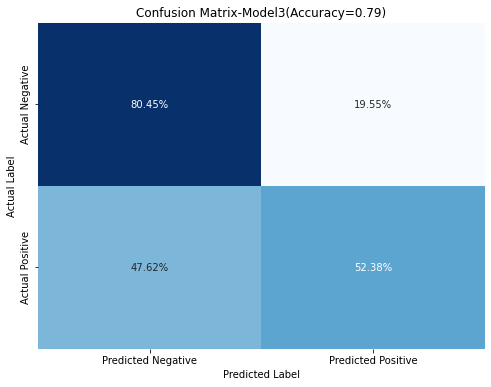
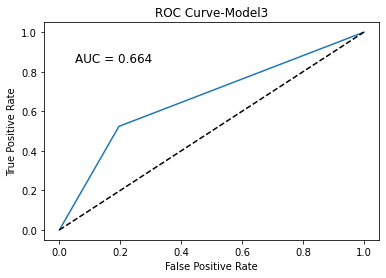
The first model is a Logistic Regression (LR) with standardization and no other feature engineering. This simple model has an accuracy of 95% and a recall of 0 on the test set by predicting every input to be negative due to the extremely minor amounts of positive samples in train and test set. The Receiver Operating Characteristic (ROC) curve, which evaluates the performance of a classifier on both negatives and positives, is a diagonal line and Area Under Curve (AUC) is 0.5, the same as a random classifier. This result suggest that accuracy alone is by no mean an effective metric since a predictor that never predict positive is of no value to QWE. As is analyzed in part 1, the most valuable predictor for QWE is one that gives the optimal combination of accuracy and recall regarding the current business circumstance. However, the real business circumstance frequently changes over time and the optimal combination of the two metrics would vary with the change. Thus, an overall index that takes accuracy and recall into consideration is required to identify the best model in general tasks. Here the AUC is chosen to represent the balanced performance of the predictor model.

With the evaluating metrics determined as AUC, corresponding methods to tackle imbalanced data problem are needed. Here, the weighted loss method is chosen instead of oversampling or under-sampling. This method penalizes misclassifying a positive sample more than misclassifying a negative sample by assigning weight to the loss function during training. By applying the weighted loss method to LR, the second model, which is one of the baseline models, is built. As is shown below, this model has an overall accuracy of 0.78 and recall of 0.52. The AUC is 0.66, representing the performance of basic structure without much feature engineering. It’s worth mentioning that this rather raw model could serve as an explanatory model, since its coefficients are instructive on the overall direction of QWE service improvement, which will be discussed in next part.

**Figure 5**. Confusion matrix and ROC of Model 1

Then, to further elevate the performance of this baseline method, Recursive Feature Elimination (RFE) is applied in Model 3 to identify and drop the redundant features. Combining the result of correlation analysis, *Customer\_Age\_inmonths, CHI\_M0, CHI\_M01, Support\_M0, Support\_M01, DaysLastLogin\_M01* are chosen as the features. Comparing with Model 2, Model 3 has better accuracy and thus higher AUC.

**Figure 6**. Confusion matrix and ROC of Model 2

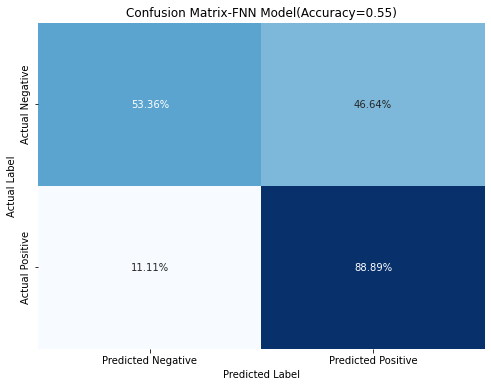
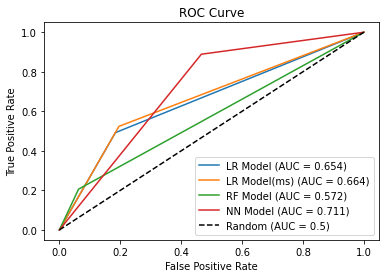
Above are the baseline chosen for this task. Besides LR models, Random Forest model is also used in experiments as a representative machine learning model, but its performance was even worse than LR models. In conclusion, traditional machine learning models do not offer an ideal solution to this problem but a baseline due to their rather simple structure and the complexity of the task. Meanwhile, Model 2 is valuable in later strategy designs and Model 3 suggests that feature selection is possible to be effective in later models.

1. Effective Predictor: Deep Learning Models

In order to solve the complicated classification task, a Feedforward Neural Network (FNN) is a strong candidate. From the experiments of machine learning models, weighted loss and feature selection are decided to be inherited. The model is designed as follows. First, to achieve the same class weight effect as in logistic regression, the binary cross entropy loss function that is used in FNN model for binary classification tasks should be redefined by assigning weight to positives, to be more specific,

,

where is the true value and is the output value before sigmoid activating. Second, to reduce the redundancy of features to a low level, L2 regularization is employed and the parameter was set to 0.025. At last, to avoid overfitting and capturing noises into the model, an early stopping with patience equals to 12 is applied. By tuning other hyperparameters, this FNN model could reach an AUC of 0.71, largely better than the baselines. The performance of this model and a thorough comparison are visualized below.

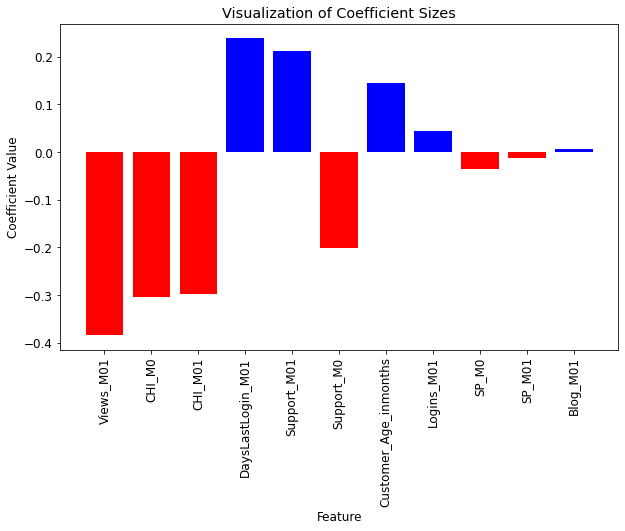
 

**Figure 7**. Confusion matrix of FNN Model and its ROC compared with others

As is shown above, this FNN model is the model with best performance so far and considered an effective predictor of churning customers in this task. Furthermore, the weight parameter, the most important parameter in this model, is capable of controlling the balance between accuracy and recall. This function is essential to real business application, which will be discussed in next part.

1. **Results and Insights**
2. Overall Service Improvement based on Logistic Regression Coefficients

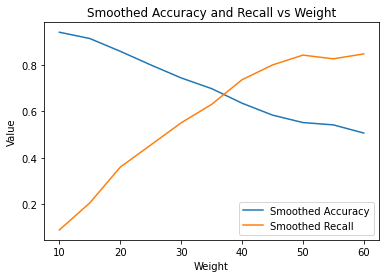
The Model 2 coefficients serve as an instruction on how to improve their overall service quality. The coefficient size plot below illustrates the rank of feature importance. To summarize, in order to retain customer in the general aspect, QWE should priorly raise customer happiness index, encourage customers to post views, make customers login more frequently and control the incremental support times under a low level. Concretely, QWE could provide benefits like tokens or gifts for customers who posted a certain amount of views as incentives. Also, perfecting the service system helps control the increase of support times, and raise customer happiness index eventually.



**Figure 8**. Feature coefficent sizes visualization based on Model 1

1. Adjustable Predictor

The preliminary analysis in part 1 concludes that the optimal combination of accuracy and recall is subject to the relation between proactive service improvement cost and customer churn loss, and the ratio of churning customers. Consequently, the predictor model is favored to be adjustable so as to cope with all situations. The FNN model in last part is an ideal candidate. As is shown below, when assigned with different weight, it performs accordingly on accuracy and recall, and the model works well with weight ranging from 1:10 to 1:50.

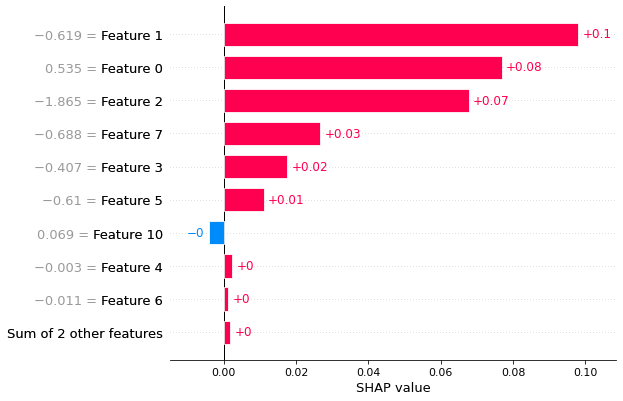


**Figure 9**. Smoothed accuracy and recall vs weight based on FNN model

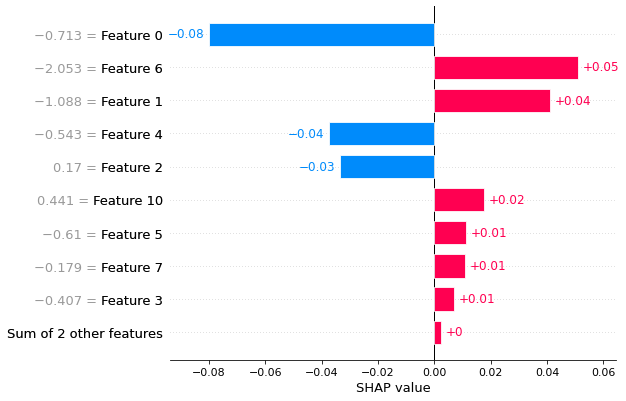
If QWE is facing a business situation that values high accuracy over high recall, a low weight ratio is suitable and the opposite otherwise. For example, consider this setting again, the loss of a churn customer is 20 times more than the customer-saving cost, and 10 percent of customers will churn. In this case, and , the total cost should be . Then the best weight to assign is 1:40 so that the overall cost is approximately . Consider another case, where all settings remain the same but , then the total cost should be , so that accuracy dominates over recall. Therefore, the best weight to assign is 1:10, and the corresponding cost is approximately .

1. Customized Retention Plan based on Shapley Additive Explanations (SHAP)

Despite that a general direction could be determined in how to improve QWE’s service quality in part (a), the specific cause of every churning customer may vary. QWE managers hope that for every customer predicted by the FNN model to churn, a customized retention plan could be designed according to the customer’s features, which is available through the help of SHAP.



**Figure 10**. SHAP analysis of customer No.74



**Figure 11**. SHAP analysis of customer No.43

Take customer No.74 and No.43 as examples. No.74 is a typical churning customer that basically all its features tells that the customer wants to unsubscribe, and the first few factors are the happiness index, customer service age and incremental login frequency. This means if QWE is to retain customer No.74, it will have to raise the customer’s satisfaction, motivate the customer to stay longer and login more frequently. While No.43 is an untypical customer that some features strongly suggest the customer would stay. In this case, QWE should especially focus on the main factors pushing the customer to unsubscribe. For customer No.43, they’re support priority and happiness index. It can be inferred that No.43 is upset with the low support priority, so that raising priority is the key to the customer’s retention.

**Appendix:**

1. **Derivation of the total cost function**

Suppose QWE now has customers, and percent of them will churn if no or passive action is taken. Every churning customer would incur a loss of , and taking proactive actions to keep one customer would generate a cost of . QWE will spend money to save every customer predicted to churn and ignore the others. Then consider a confusion matrix with each of its segment set to , respectively representing *True Negative, False Positive, False Negative and True Positive*. Then , and .

|  |  |  |
| --- | --- | --- |
| Predicted/Actual | Predicted Negative | Predicted Positive |
| Actual Negative | (TN) | (FP) |
| Actual Positive | (FN) | (TP) |

**Figure12.** Theoretical confusion matrix

Therefore, the total cost will be . So,

The last part, which is can be seen as a constant, so the determinant is:

To be noticed, this is just a simplified setting. In reality, QWE could try to take other factors like retention success rate into consideration. All these factors could be statistically estimated with past data.

1. **Variable Table**

|  |  |  |
| --- | --- | --- |
| Variable name | Variable type | Variable explanation |
| Churn | binary | whether the customer churns in 2 months |
| Cutomer\_Age\_inmonths | continuous | time length since subscription |
| CHI\_M0 | continuous | customer happiness index |
| CHI\_M01 | continuous | incremental CHI |
| Support\_M0 | continuous | frequency of support |
| Support\_M01 | continuous | incremental support times |
| SP\_M0 | discrete | level of support priority |
| SP\_M01 | discrete | incremental level of support priority |
| Logins\_M01 | continuous | incremental login frequency |
| Blog\_M01 | continuous | incremental number of blogs written |
| Views\_M01 | continuous | incremental number of views posted |
| DaysLastLogin\_M01 | continuous | incremental interval between logins |