PREDICTING CUSTOMER SATISFACTION USING DATA MINING TECHNIQUES

[1]HAIDAR ALMOHRI, [2]MOHAMMAD JAFAR ESMAEILI

[1]Siemens EES, Kuwait
[2]University of Dayton, Dayton, Ohio, USA
E-mail: [1]Haidar.almohri@siemens.com

Abstract - Customer satisfaction is one of the most important factors in determining the success/failure of a customer-focused business. Loyal customers are the main assets for businesses to survive and succeed in today's competitive market. Although businesses spend a lot of resources to collect raw data regarding the customers' behaviors, few would succeed in converting the data to meaningful results. Rigorous and advance analysis of the collected data would be required to convert the raw data into valuable and useful results. In this work, we utilize historical data, along with machine learning technics to predict customer satisfaction and determine the key-drivers for achieving a high rate customer satisfaction. The data consists of survey responses of over 76000 customers from a well-known bank. We have used several classification models for classification purpose. The result of the models are compared and analyzed to find the advantage/disadvantage of each method. Despite the fact that more than 80% of the data was missing (due to unanswered questions by the survey takers), we achieved an overall accuracy of more than 80%.

I. INTRODUCTION

Ensuring customer satisfaction and loyalty is the vital fact for long-term corporate profits and success (Rust & Chung, 2006; White, 2010). It becomes more important if we understand that today's market is so competitive and attracting customers is a difficult task. Organizations have realized that they can't compete in the market just by reducing the price of the products/services, but they need to keep a good relation with their customers (Wadud&Sharmin). They also recognize that that it is more important and profitable to keep their current customer rather than acquiring new ones (Boulter, 2013). A loyal customer can be viewed as a treasure for a corporate that needs to remain happy and satisfied with products and services of the corporate. One of the techniques that businesses utilize to identify the users' satisfaction is using surveys. Although using surveys is a convenient method to learn about customers' expectations, it has its own challenges. One of the main challenges is selecting appropriate methods to overcome the poor quality of the collected data. Missing data and erroneous responses are two of the scenarios that could jeopardize the data analysis. For these reasons, this research methodology formed into two main phases of pre-processing and data analysis phases. In the pre-processing phase this research utilized two main techniques of down sampling and exact matrix completion using soft thresholding (Candes& Benjamin Recht, 2009). Finally, this study formed three models utilizing Linear Discriminate Analysis, Random Forest, and Soft margin SVM to predict customers' satisfaction.

II. PRE-PROCESSING

A well-known bank provided the data set. The collected data is the result of surveys collected from 76020 customers. As it is usually the case with survey-based data, this data is also incomplete. The preliminary analysis of the historical data reveals that around 90% of surveys have missing entries. The large proportion of missing data increases the significant effects of data imputation. There are initially 336 columns that are used as predictors for our predictive model. The single response is the outcome of whether a customer is satisfied (1) or not satisfied (0). To overcome the challenge of missing entries, this research utilizes following steps to prepare the data prior to building the models:

A. Data Imputation and Down Sampling

Imputation is the systematic process of replacing missing data with substituted values. There have been a large number of studies focused on developing mathematical algorithms to perform imputation as a pre-processing of the incomplete data. One of the well-known techniques that have been used and validated in this field is exact matrix completion through soft thresholding. In a study conducted by Emmanuel J. Candes and Benjamin Recht (2009), it is proved that a low-rank, nxn matrix with m observed entries could be fully recovered with high probability, by solving a convex optimization problem if the inequality 2.1 holds true:

$$m \geq Cn^{1.2} r \log n \qquad (2.1)$$

where C is a constant and r is the rank of the to-be-recovered matrix [2]. The algorithm states that for a matrix **M**, if the set of observed entries is denoted as Ω $\{(i,j) \in \Omega$ if $\mathbf{M}_{ij}$ is observed$\}$, then **M** is recovered by solving the convex optimization problem:

$$\begin{aligned} \text{minimize} \quad & ||X||_* \\ s.t \quad & X_{ij} \\ & = M_{ij} (i,j) \in \Omega \end{aligned} \qquad (2.2)$$

where **X** is the recovered matrix, and $||X||_*$ is the nuclear norm of the matrix **M** (sum of its singular values).

We applied the above algorithm using the softImpute package in R and the entire code along with the rest of the analysis accessible through GitHub (Almohri, 2016).

We then removed the columns with near zero variance as well as the columns that have a correlation value greater than 0.8 to account for multicollinearity of our model.

Also, there are number of outliers in each predictor that are removed during the pre-processing. An observation is marked as an outlier if it falls beyond 95% confidence interval of its corresponding predictor. That is, for an observation $X_i^{(p)}$ (i^{th} observation of p^{th} predictor):

$|X_i^{(p)}| \in \bar{X}^{(p)} \pm 3\sigma_X^{(p)}$ (2.3)

where $\bar{X}^{(p)}$ and $\sigma_X^{(p)}$ are the mean and variance of the p^{th} predictor.

After pre-processing, 107 columns remain as predictors in the data set that are fed to our predictive model.

One of the challenges that are encountered in this work is that the labels in the data set are imbalanced, i.e. the number of one class is highly larger than the other class. In particular, about 96% of observations belong to class 0 (satisfied customers), while the rest of the observations belong to class 1 (unsatisfied customers). The imbalanced data set can arise in many real-world problems. Examples include fraud detection, network intrusion, rare disease diagnosing, etc. (Chen, Liaw, Breiman, 2004). The main problem in modeling such population is that the general classifiers aim to minimize the classification error rate, while in such problems there should be a special attention given to the rare class. As a result, the misclassification error may be misleading. For example, if the model classifies all the observations as 0, the error rate will be 4%, which obviously carries 100% false negative error. With increasing availability of data, there has been an increasing attention to this problem in recent years (Chen, Liaw, Breiman, 2004; Chawla, 2009; Provost, 2000).

The other technique that has been used during the training process is down sampling. Down sampling is simply reducing the amount of over-represented class so that there will be a balance between the two classes, e.g. half-half. Although this technique has a drawback, which is losing a fair amount of our data, it has proved to be a very good strategy in practice. Hence, the data set is split into training (80%) and testing (20%) using down sampling.

III. PREDICTION MODELING

Three different classification algorithms are applied to the data set to find the best model that produces the best results. In this section, brief overviews of the utilized models as well as preliminary results are presented. A more in-depth discussion about the results as well as the comparison between the models is presented in the following section.

A. Linear Discriminant Analysis

Given the data $X \in R^p$, and labels Y, Linear Discriminant Analysis (LDA) is among the family of so called "generative models", which model the posterior distribution P(Y|X) by estimating joint distribution P(X,Y) (in contrast of "discriminative models" such as logistic regression that model the posterior distribution P(Y|X) directly). In this case, the prior probability of $Y \in \{0,1\}$ is given by:

$Y \sim \text{Bernoulli}(\pi)$

where $\pi = P(Y=1|x)$, or simply the proportion of class 1 in the training data. The conditional probability of X given Y is assumed to follow a Gaussian distribution:

$P(X|Y) \sim N(\mu, \Sigma)$

where $\mu = \{\mu_1, \mu_2, \ldots, \mu_p\}$ is the vector of means and $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$. Note that it is assumed that the two classes share the same diagonal covariance (with different means).

The posterior probability is then calculated by

$$P(Y = 1|x) = \frac{1}{1+e^{-\beta^T x - \alpha}} \quad (3.1)$$

where

$$\beta = \Sigma^{-1}(\mu_1 - \mu_0) \quad , \quad \alpha = -\frac{1}{2}(\mu_1 - \mu_0)^T \Sigma^{-1}(\mu_1 - \mu_0) + \log\frac{\pi}{1-\pi} \quad (3.2)$$

As it appears in equation 3.2, the posterior probability is influenced by the choice of prior probability (π). Using this fact, and especially in the case of imbalanced data, one can fine-tune the prior (π), to optimize the performance on the testing data. The optimal prior probabilities are selected by cross-validation. Table 1 shows the confusion matrix for this model on the testing data set.

TABLE I
CONFUSION MATRIX FOR LDA

		Actual: 0	Actual: 1
Predicted	0	72%	27%
	1	28%	73%

B. Random Forest

Random forest is a bagging technique that is proved to be extremely powerful in both regression and classification. It randomly selects a subset of the data for n number of times to generate n decision trees. The new data point that needs to be classified is then predicted using each of n decision trees to generate n results (votes). In classification problems, the predicted result is the class that acquires highest numbers of votes. The result of applying random forest to our data set is shown in table 2.

TABLE 2
CONFUSION MATRIX FOR RANDOM FOREST

		Actual: 0	Actual: 1
Predicted	0	80%	28%
	1	20%	72%

C. Soft Margin SVM

Support Vector Machine (SVM) is a classification algorithm that is purely obtained by solving an optimization problem. Assuming that the class labels $Y \in \{-1,1\}$, SVM finds the hyperplane that best separates the two classes. This is achieved by finding the maximum margin (maximum distance between the two classes). This is called hard-margin SVM, which forces a rigid margin, i.e. a perfectly separable margin or zero error. This method works perfect for linearly separable data. But almost none of the real-world classification problems are linearly separable. Hence, allowing some data points to cross the separating hyperplane (some error) can help to build a more general model that more robustly fits the testing data. The soft-margin SVM is achieved by introducing the slack variable. The slack variable is a notion on how much we allow violation of the margin. By minimizing this variable, we can separate the data as clean as possible. If we denote the data points $(x_i, y_i)_{i=1}^n$, the distance to be minimized by w, and the slack variable by φ_i, then equation [3.3] shows the formulation of soft-margin SVM.

$$\min \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\varphi_i$$
$$s.t.\ y_i w^T x_i \geq 1 - \varphi_i, \quad \varphi_i \geq 0 \quad (3.3)$$

where the constant C is the penalty term (a.k.a regularizer), that defines how much should each violation be penalized when training the model.

Moreover, one can utilize the so-called "Kernel Trick" along with SVM. Applying Kernel is an extremely powerful technique that can transfer a non-separable data to being separable in a different (higher) dimension. The idea is to project the columns (features) of the data to a higher dimension using a mapping function Θ (.), which induces the corresponding kernel function K (Jakkula, 2006; Yu and Kim, 2012).

A soft-margin SVM using two different kernels (Radial Basis Function, and second order Polynomial) is fit to our data set. Optimal parameters including the regularizer$C$ is found using cross validation. Tables 3 and 4 summarize the result of the models.

TABLE 3
CONFUSION MATRIX FOR SVM WITH RBF KERNEL

		Actual: 0	Actual: 1
Predicted	0	77%	36%
	1	23%	64%

TABLE 4
CONFUSION MATRIX FOR SVM WITH POLYNOMIAL KERNEL

		Actual: 0	Actual: 1
Predicted	0	80%	28%
	1	20%	72%

IV. ANALYSIS OF THE RESULTS

Table 5 provides detailed information about the four different classifiers that were used and mentioned in previous section.

One of the most widely used methods for inspecting the performance of binary classifier is using Receiving Operating Characteristic (ROC) curves. It is basically a way to compare the rate at which the classifier makes true positive as well as false positive predictions. In particular, area under the curve (AUC) is used to compare the mentioned classifiers. Figure 1 shows the ROC curve and table 5 illustrates the AUC obtained for each classifier.

TABLE 5
DETAILED RESULTS FOR THE CLASSIFIERS

Classifier	Accuracy (95% CI)	Sensitivity (%)	Specificity (%)	Brier Score
LDA	(82.2, 83.4)	83.7	62.2	0.17
SVM (RBF Kernel)	(76.2, 77.6)	77.5	63.7	0.18
SVM (Polynomial)	(71.4, 72.8)	72.2	68.2	0.18
Random Forest	(79.7, 80.9)	80.7	71.8	0.16

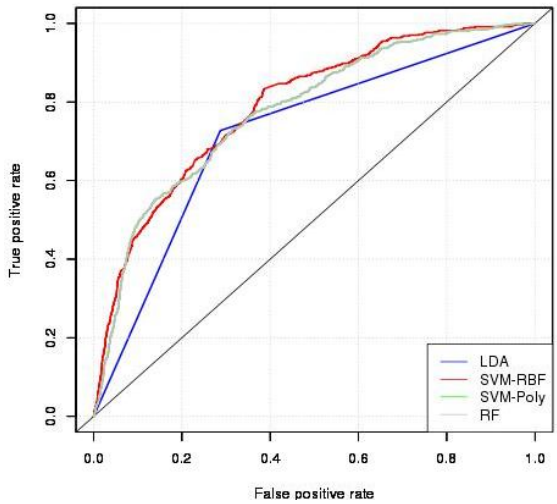


Fig. 1ROC curve for the classifiers

TABLE 6
AUC FOR THE CLASSIFIERS

Classifier	AUC
LDA	0.72
SVM (RBF Kernel)	0.79
SVM (Polynomial)	0.77
Random Forest	0.77

DISCUSSION

Assessing the accuracy and comparing the classifiers is a very widespread problem (Labatut and Cherifi, 2012). Simply looking at the number of correctly predicted values doesn't provide a strong baseline for assessing the performance of a classifier, especially in the case of heavily unbalanced data. Although looking at the sensitivity, specificity, and AUC provides some good insights about the performance of classifiers, however, these methods fail to consider the assertion of class probabilities or ranking. There are different proper scoring functions that measure the accuracy based on class probabilities. Among all, Brier score is widely used for this purpose. Brier score is a simple quadratic scoring function, which calculates the mean square error of a probability forecast. The most commonly used formula that is used to calculate the Brier score is:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2 \quad (5.1)$$

Where p_i is the probability that the model is produced for observation i, and o_i is the observed label. As a result, a lower Brier score for a classifier shows a higher performance. Combining the absolute prediction results with some proper scoring rule, e.g. Brier Score can provide more insights when assessing classifiers. Based on the results mentioned in previous section, LDA produced the highest overall accuracy (with 95% Confidence Interval of (82.2, 83.4)). However, most of its gain is obtained in predicting the true positive class, which represents predicting the satisfied customers in this study. On the other hand, it has the least accuracy in predicting the negative class (unsatisfied customers) as well as the least AUC. It also had the second lowest Brier score. Both SVM models (with RBF and second order polynomial kernels) produced similar results, with the difference that using RBF kernel results in higher sensitivity (77.5%), whereas second order polynomial provides higher specificity (68.2%). SVM with RBF kernel produced the highest AUC. They both had the highest Brier score. Random forest performed quite well on both sensitivity and specificity (second highest sensitivity and the highest specificity). It provides a good balance between the two prediction classes. In general, random forest is a very strong tool in situations where only the prediction and not explanation of the result is required. Random forest also had the lowest Brier score.

CONCLUSION

Four different models for predicting customer satisfaction using a highly incomplete and imbalanced data are presented. Ignoring some deviation, all models performed quite well on the testing data. Since each classifier performed differently on predicting positive or negative classes, the best way to select the final classifier is to provide a cost for misprediction. For example, if the cost of classifying a satisfied customer as unsatisfied is proved to be higher than classifying an unsatisfied customer to satisfied, then one would select LDA, which has the highest sensitivity. Otherwise, random forest can be selected.

The ROC curve can provide a baseline for the business operators to find the tradeoff between the sensitivity and specificity on different cut-offs. The proper cut-off point can be achieved according to the need of the business (the cost of type I and type II errors).

REFERENCES

[1] Almohri, H., Utilizing R programming to predict customers' satisfactions utilizing advanced techniques, (2016), GitHub repository https://gist.github.com/mesmaeili1/48cb2aa5eb91b6704aac7b30e4bff831.

[2] Candès, E.J. and Recht, B., 2009. Exact matrix completion via convex optimization. Foundations of Computational mathematics, 9(6), pp.717-772.

[3] Chawla, N.V., 2009. Data mining for imbalanced data sets: An overview. InData mining and knowledge discovery handbook (pp. 875-886). Springer US.

[4] Chen, C., Liaw, A. and Breiman, L., 2004. Using random forest to learn imbalanced data. University of California, Berkeley.

[5] Jakkula, V., 2006. Tutorial on support vector machine (svm). School of EECS, Washington State University.

[6] Labatut, V. and Cherifi, H., 2012. Accuracy measures for the comparison of classifiers. arXiv preprint arXiv:1207.3790.

[7] Provost, F., 2000, July. Machine learning from imbalanced data sets 101. InProceedings of the AAAI'2000 workshop on imbalanced data sets (pp. 1-3).

[8] Yu, H. and Kim, S., 2012. Svm tutorial—classification, regression and ranking. In Handbook of Natural computing (pp. 479-506). Springer Berlin Heidelberg.

[9] White, C. J. 2010. The impact of emotions on service quality, satisfaction, and positive word-of-mouth intentions over time. Journal of Marketing Management, 26(5-6), 381-394.

[10] Rust, Roland T., and Tuck Siong Chung. Marketing models of service and relationships. Marketing science 25.6 (2006): 560-580.

[11] Reinartz, Werner, Manfred Krafft, and Wayne D. Hoyer. The customer relationship management process: Its measurement and impact on performance. Journal of marketing research41.3 (2004)

[12] Wadud, Sharmin. Customer satisfaction in business: A case study of Moon travel ltd. (2013).

★★★