CKME 136 Data Analytics: Capstone Course

Final Report

Chang School of Continuing Education, Ryerson University

Title: The Effects of Perceived Quality and Satisfaction on Hospital Rating

Jae Yong (Francisco) Lee

#500 883 333

Date Submitted: December 3rd, 2018

Supervisor: Dr. Karim El Mokhtari, [elmkarim@ryerson.ca](mailto:elmkarim@ryerson.ca)

**Introduction**

There is complexity in evaluating the efficacy of a health care system due to a combination of scientific, technological, economic, and epidemiological imperatives [1]. As such, the common proxy for quality of hospital care is the patients’ ratings of their hospital [2]. The perceived quality and satisfaction of patients is of particular importance as it can be a salient indicator used to draw detailed contextual descriptions of hospital’s different components to identify patterns that may have an impact on other outcomes of a health care system. In addition, the findings of these comparative analyses can recommend policies that are aimed at system-level improvements. I will evaluate the official hospital compare datasets provided by Center for Medicare and Medicaid (CMS) for hospital quality comparison [3]. Classification methods will be used to predict the overall hospital rating and identify the characteristics of hospitals. For predictive modelling, three classification algorithms will be explored: (1) decision tree, (2) Naïve Bayes, and (3) logistic regression. These methods will be implemented using a combination of Weka and R.

**Literature Review**

The quality of health care is hard to quantify as it is contingent on several factors. Based on the literature review, there is a lack of exploration on the hospital performance on measurements of patients’ ratings and performance with respect to one characteristic of hospitals (e.g., mortality national comparison) related to performance with respect to another characteristic (e.g., readmission national comparison). In addition, the quality metrics for characterizing a hospital of high quality has yet to be discussed. For example, do patients who receive care in hospitals with key characteristics (not-for-profit hospital, below national average timeliness, etc.) report better experiences than patients in hospitals without these characteristics? DeLancey et al. in 2017 published their results on the associations between hospital characteristics, measure reporting, and the CMS Overall Hospital Quality Star Ratings [4]. Based on this article, hospitals with star ratings between 4 and 5 were considered to have high quality while hospitals with 1 to 3 stars were considered to have low quality. The principal investigators have concluded that smaller hospitals more frequently achieved a high star rating compared with larger hospitals.

National policies to improve health care quality have largely focused on clinical provider outcomes and payment reform; however, the study conducted by Tsai et al. in 2015 revealed the association between hospital leadership and quality [5]. Also, the paper outlines the increase in quality of care with hospitals practicing more effective management. In particular, the hospital ownership type determined the management style which validates the relevance and inclusion of the Hospital Ownership attribute in the analysis. In addition, Bloom and his team explored whether different management style affected hospital performance [6]. They published that higher competition results in higher management quality and improved hospital performance.

With regards to the meaningful use of electronic health records (EHRs), it is a challenge to validate the relevance as the external rules, regulations and pressures have influenced the credibility of such a metric. The recent passage of the American Recovery and Reinvestment Act (ARRA) of 2009, which includes the Health Information Technology for Economic and Clinical Health (HITECH) Act, made available over $20 billion dollars for health care practitioners who become “meaningful users” of health IT. Thus, ARRA introduces the single largest financial incentive ever to facilitate electronic health record (EHR) implementation. Recent studies reveal strong evidence that supports the superficial validation and implementation of the meaningful use of EHRs criteria, as such, the attribute regarding meaningful use of EHRs will be eliminated [7].

**Dataset**

The dataset used in this analysis is sourced from the Hospital ratings dataset from Kaggle [8]. This is derived from the official dataset used on Medicare.gov (<https://www.medicare.gov/>) for hospital quality comparison.

The dataset consisted of 4812 observations and 28 attributes. The nominal variables include Hospital Name, Address, City, State, County Name, Hospital Type, Hospital Ownership, Emergency Services, Meets criteria for meaningful use of EHRs, and the following variables had themselves and the respective footnotes: Hospital Overall Rating, Mortality National Comparison, Safety of Care National Comparison, Readmission National Comparison, Patient Experience National Comparison, Effectiveness of Care National Comparison, Timeliness of Care National Comparison, and Efficient Use of Medical Imaging National Comparison. The criteria to include and/or exclude certain attributes will be further discussed in the Approach section, in particular with reference to redundant attributes.

**Approach**

**Step 1: Data Preparation**

The detailed approach accompanied by the respective programming can be found on the GitHub page (<https://github.com/fleejy/ckme136-capstone/>).

The dataset used in this analysis consisted of 4812 observations and 28 attributes. Based on literature review, the *Hospital overall rating* variable which act as the key prediction variable was converted to a nominal variable. Other nominal variables include: Hospital Name, Address, City, State, County Name, Hospital Type, Hospital Ownership, Emergency Services, Meets criteria for meaningful use of EHRs, Hospital overall rating footnote, Mortality national comparison, Mortality national comparison footnote, Safety of care national comparison, Safety of care national comparison footnote, Readmission national comparison, Readmission national comparison footnote, Patient experience national comparison, Patient experience national comparison footnote, Effectiveness of care national comparison, Effectiveness of care national comparison footnote, Timeliness of care national comparison, Timeliness of care national comparison footnote, Efficient use of medical imaging national comparison, and Efficient use of medical imaging national comparison footnote.

The numeric variables consist of Provider ID, ZIP Code, and Phone Number. They were of no significance to the purpose of this analysis; therefore, the overall dataset is qualitative such that summary statistics was not appropriate. Consequently, Principal Component Analysis (PCA) could not be conducted for dimensionality reduction; however, Multiple Factor Analysis (MFA) which is dedicated to datasets where variables are structured into groups was considered [9]. However, once the missing data and outlier treatment was conducted the dimensionality of the dataset was reduced to manageable size.

From the Exploratory Data Analysis (EDA), the primary objective was to characterize hospital overall quality rating and to find a typology of the hospital overall quality rating. Attributes that did not contribute to the modelling purpose were eliminated (i.e., Provider ID, Hospital Name, Address, City, State, ZIP Code, County Name, Phone Number, and all attributes’ footnotes) [Figure 1.3].

In addition, outlier detection was not suitable for a categorical dataset; therefore, chi-squared test was conducted for all attributes. Chi-squared test was chosen because it is able to determine if the attributes are dependent or not, given that both response and predictors are categorical variables (i.e., hospital overall rating and all other attributes) [Figure 1.2]. For all attributes except Emergency Services attribute, the null hypotheses were rejected because the p-values of the observed H statistics were smaller than 0.05. This indicates that the attributes differ in their overall hospital rating, suggesting a significant effect in the overall rating. As such, these attributes were included in the analysis. On the other hand, Emergency Services’ null hypothesis was retained because the p-value of the observed H statistic was larger than 0.05 which suggest that the attribute was not significant in its effect in relation to the overall hospital rating. As such, the Emergency Services attribute was removed.

With regards to missing data, the Hospital Type required careful analysis with respect to the national comparison attributes. It is worth noting that all instances of the Critical Access Hospitals and Children’s Hospital had missing data for the national comparison attributes.

This may be explained through the nature of these hospital types. Critical Access Hospital (CAH) is a designation given to eligible rural hospitals by the Centers for Medicare and Medicaid Services (CMS). Furthermore, any rural hospitals that are CAH designated receive certain benefits, such as cost-based reimbursement for Medicare services which help to reduce their financial vulnerability while improving access to healthcare by keeping essential services in rural communities. As such, it is expected any tracking of information was not available; therefore, any hospital types of CAH was removed. For Children’s Hospital, the confidentiality clause of minor prevents the collection of any metrics required for this analysis. As such, Children’s Hospital instances were further removed.

The complexity Children’s Hospital and CAH are both instance specific such that they are deemed unsuitable for the purpose of this analysis. Furthermore, the uniqueness of both Children’s hospitals and CAH require dedicated studies. Given these characteristic and elimination of the two classes, the Hospital Type attribute was additionally removed.

After the data was cleaned, there was a total of 2200 observations and 9 attributes and the following abbreviations were used for attribute names:

|  |  |
| --- | --- |
| **Original Attribute Name** | **Abbreviated Attribute Name** |
| Hospital Ownership | h\_ownership |
| Hospital Overall Rating | h\_rating |
| Mortality National Comparison | h\_mortality |
| Safety of Care National Comparison | h\_soc |
| Readmission National Comparison | h\_ra |
| Patient Experience National Comparison | h\_pex |
| Effectiveness of Care National Comparison | h\_eoc |
| Timeliness of Care National Comparison | h\_toc |
| Efficient Use of Medical Imaging National Comparison | h\_imaging |

The dataset showed significant imbalanced distribution with 1588 False for not HHQ (i.e., LHQ) and 612 True for HHQ. I used *ClassBalancer* in Weka filter to equalize the number of False to make the ratio of the two classes equal. *ClassBalancer* reweighs the instances in the data so that each class has the same total weight, while maintaining the total sum of weights across all instances. Therefore, the final number of sample was 2200 (i.e., 1100 for both HHQ and not HHQ while maintaining 612 True for HHQ and 1588 False for HHQ). All the attributes were statistically significant, and with no missing values in the dataset, the dataset was considered clean [Figure 1.3].

**Step 2: Predictive Modelling**

## Decision Tree

The first classification algorithm used for predictive modelling was the decision tree model. With the prepared dataset, initial assessment of the models displayed high accuracy rates around 90% [Figure 2.1.1]. The primary metric of recall was significantly lower compared to the overall accuracy rate. From the initial assessment, I concluded that the model was biased toward the false instances as a consequence of the imbalance in distribution of the class attribute. The imbalance resulted in high true negative rate at the cost of low true positive rate.

Two approaches were explored to resolve the imbalance issue. The first approach was to over-sample true instances, and the second approach was to down-sample false instances. The over-sampling option was discarded because there was large enough dataset to down-sample from and oversampling generally makes overfitting more likely. In order to overcome the imbalance, *ClassBalancer* in Weka filter was used to equalize the number of False to make the ratio of the two classes equal. Proceeding with the class-balancing approach, the instances were reweighed such that each class had the same total weight, while maintaining the total sum of weights across all instances.

With the balanced dataset, J48 Classifier from Weka was used to construct decision tree model. Within different parameters present in Weka, *minNumObj* parameter was used to obtain models with different tree sizes. Summary of key metrics on [Figure 2.1.2 and Figure 2.1.3] shows that class-balancing significantly enhanced the recall of the model. The model with the second smallest tree size among six models with various tree sizes was chosen, because the model displayed highest recall, maintained acceptable rate of accuracy, and was most comprehensible. The recall of the final model was 83.20% and the accuracy was 82.64%.

The final decision tree model consisted of nineteen leaf nodes with three attributes as decision rules [Figure 2.1.4]. The three attributes that played vital role in the decision-making process were Safety of Care, Readmission, and Patient Experience. Hospitals with Safety of Care above national average, hospitals with Readmission below national average, and hospitals with Patient Experience above national average were more likely to have HHQ.

## Naive Bayes

The second classification algorithm used for predictive modelling was the Naive Bayes model, which assumes that the input values are nominal (numerical inputs are supported by assuming a distribution). Naive Bayes uses a simple implementation of Bayes Theorem (hence naive) where the prior probability for each class is calculated from the training data and assumed to be independent of each other (conditionally independent), which is an unrealistic assumption because we expect the variables to interact and be dependent. However, such an assumption allows for faster and easier calculation of the probabilities.

With the initial dataset, the assessment of the models displayed high accuracy rates around 87% [Figure 3.1]; however, the recall and accuracy varied depending on the type of test option which will be explained in a following section. This accuracy is promising because Naive Bayes model is can handle categorical inputs by default, as such the accuracy is representative of the model’s strength.

For this data set, Naive Bayes is suitable. This is due to the strength of Naive Bayes in handling nominal attributes. If the attributes numeric, the model would require treatments such as (1) discretizing the numeric variables or (2) using probability density functions. Discretizing the continuous variable would be a simple solution; however, it is subjective which causes loss of information, but it is used as a quick way to prepare data before applying Naive Bayes classification.

By default, a Gaussian distribution is assumed for each numerical attribute in Weka. In order to handle numerical attributes, the numerical attributes are automatically converted to nominal attributes with the ***useSupervisedDiscretization*** parameter set as TRUE. The particular implementation of Naive Bayes algorithm would be changed to use a kernel estimator with the ***useKernelEstimator*** argument which can be used to better match the actual distribution of the attributes in datasets. It is important to note that ***useKernelEstimator*** and ***useSupervisedDiscretization*** parameters are mutually exclusive.

After conducting various test options on the class-balanced dataset, the test option which produced the best recall rate and accuracy was 70% split.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test Option** | **Recall (T)** | | | **Accuracy** |
| 70% split | | 89.6% | 85.34% | |
| Use training set | | 87.6% | 84.59% | |
| 10-fold CV | | 87.6% | 84.47% | |
| 3-fold CV | | 87.4% | 84.33% | |

*Figure 3.4.2*

Going back to data preparation, in order to treat for imbalanced class distribution, I utilized class-balancing which resulted in limited number of samples; therefore, I ruled out using the training set as the test option due to its small sample size. More importantly, training set was not used because it produced generalization error (i.e., using only one data set, the model could achieve high accuracy by simply learning this particular set, but not the general concept). In accordance with the former reasons, the percentage splits were ruled out due to generalization error [10]. In addition, the use of training/test sets and cross-validation are conceptually equivalent. Cross-validation simply takes a more rigorous approach by averaging over the entire data set. Consequently, the n-fold cross-validation (CV) was chosen [11].

Between the 10-fold and 3-fold CV, the 10-fold CV was chosen over 3-fold because a larger *n*-fold has less bias towards overestimating the true expected error (i.e., training folds closer to the total dataset), but higher variance and running time (i.e., approaches the limit case of leave-one-out CV) [12]. In addition, higher *n* gives you more samples to estimate a more accurate confidence interval on your estimate [13].

Based on the classification using Weka, I have found that the 10-fold cross validation was most suitable with a recall rate of 87.6% and accuracy of 84.47%.

## Logistic Regression

Besides Decision Tree and Naive Bayes, logistic regression was chosen as the third classification algorithms to build the predictive model. Logistic regression converts binary classification problems into linear regression ones. That is, by means of proper transformation, it is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables. In our case, it predicts the probability of hospital overall quality for a hospital by fitting data to a logit function with factors affecting hospitals’ ratings.

I have found that for logistic regression in particular, there was of little benefit to creating a balanced sample. Compared to balancing the class, using the original datasets containing all the information available improves classifier performance. Thus, I fitted the baseline model in Weka with 70% training set from the original dataset, with outliers removed and attributes with little contribution to our model construction, such as State, Phone and ZIP Code, dropped.

From the model summary [Figure 2.3.1], we can see that the estimated coefficients of *Patient Experience National Comparison*, *Mortality National Comparison*, and *Readmission National Comparison* are statistically significant variables suggesting strong associations with the probability of high hospital quality. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable. For example, the positive coefficient for *Readmission* predictor suggests that all other variables being equal, the hospitals who have *Readmission* predictor are more likely to be a high quality hospital.

Based on most relevant attributes, the final model is built with all estimated coefficients being statistically significant. As shown in [Figure 2.3.3], the signs of the estimated coefficients suggest that being a customer with international plan will increase the probability of churn, while being a voicemail plan subscriber will decrease the probability of churn. Units increase in Day, evening, night or international minutes and more customer service calls will also increase the probability of churn. The AIC of our final model, an analogous metric of adjusted R-square in logistic regression, reduces compared to our initial model, indicating this model is a better fit with less relative information lost. Also the 86.6% accuracy seems to be a good result, although it is dependent on the manual split of the data we made earlier. The results from 10-folds cross-validation for the final model shows similar accuracy. The detailed accuracy by class and confusion matrix are shown as [Figure 2.3.5].

When selecting the best performing algorithm, we want the algorithm which maximizes the true positives and true negatives because we have imbalanced class distribution. If we only look at accuracy, logistic regression seems to be the best model to make the predictions. However, accuracy works best if true positives and true negatives have similar cost. In our case, the cost of true positives and true negatives are very different, so it’s better to look at both Precision and Recall since we are emphasizing on customers who are more likely to stop their services with the company.

During model evaluation of Logistic Regression on the 30% testing dataset, we noticed that the True Positive Rate, False Positive Rate and Recall are relatively low even if we tried to set different classification thresholds (from 0.1 to 0.5), implying our regression model did not predict well on positive observations, which are customers who have high probability of churn. Logistic regression run on down-sampled dataset significantly improves precision and recall with overall accuracy being compromised. This may be caused by the impact of imbalanced data on the estimated coefficient of the intercept, which eventually affects the log odds of Churn and the testing results. Another informative technique we used to see our classifier performance is ROC curve. ROC is a plot of True Positive Rate against False Positive Rate on every possible classification threshold. In other words, ROC visualizes sensitivity versus specificity for all possible cut-off classification probability values, whereas the classification table only represents error rate for a single threshold. In our final model of logistic regression, we can see the ROC curve [Figure 2.3.4] is in the upper left corner of the plot, indicating our model does a good job in separating classes. And Area Under Curve (AUC) is around 0.8, suggesting a strong prediction power of the model.

The conclusions we drew from Logistic Regression is consistent with the analysis we did from the Decision Tree -- international plan subscribers, customers with more daytime usage and customers who make more customer service calls are more likely to churn. However, Decision Tree provides higher precision and recall than Logistic Regression. And as discussed in the previous section, Naive Bayes is not suitable for our dataset. Thus we select Decision Tree as the best performing algorithm in this case.

The comparison is shown as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision (True) | Recall (True) | ROC Area |
| Decision Tree | 81.24% | 80.13% | 83.08% |  |
| Logistic Regression (on full dataset with classification threshold = 0.5) | 86.6% | 61.2% | 21.9% | 0.814 |
| Logistic Regression (on down-sampling dataset) | 78.1% | 79.4% | 75.9% | 0.832 |

Step 3: <Post-predictive Analysis>

1. Clustering analysis vs Association rule?
   1. Clustering requires quantitative whereas, association requires categorical attributes; therefore, association will be tested.

# Post-predictive Analysis

For the post-predictive analysis, I used clustering analysis because majority of the attributes were qualitative. The association rule was not examined, because the type of attributes was not suitable for the optimum performance of the model (i.e., association rule requires categorical attributes).

To successfully identify the characteristics of churning customers, we first filtered not-churning customers out of the dataset. Then we applied SimpleKMeans Clusterer to the filtered dataset with different numbers of clusters.

We examined six different models with number of clusters ranging from two to seven. As shown in [Figure 4.1.1], the sum of squared errors decreased the most when the number of cluster was increased from two to three. We decided that the best number of clusters is three because adding another cluster beyond this point did not give significantly better modelling of the data. Three centroids of the clusters from the model are displayed in [Figure 4.1.2].

On average, cluster 1 showed higher customer service calls with lower number of calling minutes, cluster 2 showed higher number of calling minutes, and cluster 3 showed higher number of voicemail messages.

Recommendations:

* Cluster 1: One possible explanation behind this cluster is that there are customers experiencing problems on calling, so they tend to have lower usage minutes and higher number of customer service calls. We recommend further technical analysis on this cohort and surveying customers on their experience with the customer service call centre.
* Cluster 2: We suspect customers that this cluster of customers are leaving because their total bill amount is too high. We advise implementing discounted plans for heavy users.
* Cluster 3: The reason behind customers churning could be current voicemail service being too expensive, or the lack of functionality and quality. We recommend auditing the current voicemail plan.

Key traits of each clusters are graphically displayed in [Figure 4.1.3].

**Results**

**Explain your results here. Consider that you need to communicate your results to executives in an organization. For example:**

* + - 1. Insert tables and/or charts showing the results
      2. Write description of the tables and charts, such that they show the usefulness for an organization
      3. Identify the evaluation measures, such as accuracy, precision, recall, etc.

Once the analysis is complete, the conclusion will summarize the results and recommendations will be given where detailed contextual descriptions of hospital’s different components may be drawn to identify patterns that may have an impact on other outcomes of a health care system. In addition, recommendations will be made based on the findings of the comparative analyses that are aimed at system-level improvements.

# Conclusion and Recommendations

Having the ability to predict and act on insights for customer retention is a priority for any successful company. This can be achieved through data analytics and optimizing workflows based on those findings. Through our analyses, we have found several key points that may be of significant contribution to further developing such an ability to understand our customers leading to robust business execution.

After initial analysis, the data was prepared for our investigation. The eight non-relevant attributes were removed from 21 total attributes resulting in 13 attributes and 1 class attribute (churn). In addition, the dataset showed significant imbalanced distribution with 2850 False for not churn and 483 True for churn. We used downsampling in Weka filter to bring down the number of False to make the ratio of the two classes equal. Therefore, the final number of sample was 922 (i.e., 461 True for churn and 461 False for churn). The attributes were normal and without missing values, at which point, we considered the dataset clean.

For our predictive modelling, three types of classifiers were tested: (1) Decision Tree, (2) Naive Bayes, and (3) Logistic Regression. The explanation on criterion for ruling out the classifier is detailed in the classifier’s respective section. The Naive Bayes classifier was ruled out due to its assumptions which were not suitable for our analysis. The Logistic Regression classifier was ruled out because the accuracy of the model was high with the recall being disproportionately low. The Decision Tree was chosen as the most suitable classifier for our analysis, because the model displayed highest recall, maintained acceptable rate of accuracy, and was most comprehensible.

For the decision tree classifier, we used recall (i.e., recall on true instances) and accuracy (i.e., proportion of correct classifications among all classifications) which assigns equal cost to false positives and false negatives. When the data set is imbalanced, the accuracy is a poor measure; however, our data set was balanced. Thus, the problem was eliminated. The recall of the final model was 83.08% and the accuracy was 81.24%.

For the post-predictive analysis, we used clustering analysis because majority of the attributes were quantitative. Using the association rule would result in suboptimal performance of the model (i.e., association rule requires categorical attributes). Using k-means clustering algorithm, we examined six different models with number of clusters ranging from two to seven. When the number of cluster was increased from two to three, the maximum decrease in sum of squared errors was observed. The best number of clusters is three, because adding another cluster did not significantly contribute to a better modelling of the data.

From the decision tree, we have identified three key attributes: (1) subscription to international plan, (2) number of calls to the customer service centre, and (3) usage of day minutes. These attributes should be carefully monitored as they are the key attributes which lead to the churning of customers. Therefore, we recommend monitoring customers on these attributes to identify their likelihood of churning. This insight is of uttermost importance to the prosperity of the company, because it is much less expensive to retain existing customers than it is to acquire new customers—earning business from new customers means working leads all the way through the sales funnel, utilizing the company’s marketing and sales resources throughout the process. Customer retention is more cost-effective as we have already earned the trust and loyalty of our existing customers; therefore, having insight to predict churn is critical to our company’s success.

Based on our post-predictive analysis, our company must pay careful attention to customers who portray a higher customer service calls with lower number of calling minutes, a higher number of calling minutes, and a higher number of voicemail messages are more likely to churn. As such, developing strategies to ensure their continued loyalty is paramount to this report.

We recommend that the company investigate, for customers with more frequent calls to the customer service centre with low calling minutes, whether they are experiencing technical difficulties or if they are unable to make calls at all. If it is the former, the company must act promptly to resolve the technical issues. If it is the latter, we recommend conducting further technical analysis on this cohort and collect surveys from customers on their experience with the customer service call centre or their subscribed services. For customers who have higher call usage, we recommend implementing tiered discounted plans (e.g., unlimited usage at a static price) for these heavy users. For customers who have higher number of voicemail messages, the reason may be of a twofold nature: (1) the current voicemail service is costly or (2) the lack of voicemail service’s functionality and quality. As such, we recommend auditing the current voicemail plan and reassessing the functionality and quality of the service.

* Prediction Model showed that these attributes lead to churn:
  + Number of calls to the customer service centre
  + Subscription to the international plan
  + Usage of day minutes
* Post-predictive Model demonstrated that churned customers had these attributes:
  + Higher customer service calls with lower number of calling minutes
  + Higher number of calling minutes
  + Higher number of voicemail messages

The findings from our prediction model and post-predictive analysis differ in that prediction model suggests which key attributes our company must monitor to track potential customers who churn and our post-predictive analysis suggests characteristics of customers who have churned; therefore, providing insight into which business strategies must be employed to retain customers.

# 

**Conclusions**

**Give a short summary (one to two paragraphs) of your analysis and conclude the discussion by defining the usefulness of your analysis.**

Our company has provided excellent telecommunication service to all our customers. Customer attrition (hereinafter interchangeably referred to as customer churn) occurs when customers stop doing business with our company. Unfortunately, we do not have any prediction model to gain insight on why certain customers discontinue our services nor a metric to track and predict customer churn. Given such metric’s saliency, we have utilized predictive modelling and post-prediction analyses to gain insight into how we may be able to predict which customers would churn. Based on decision tree classifier, the recall on true instances, hereinafter simply referred to as recall, was 83.08% and the accuracy was 81.24% (accuracy defined as the proportion of correct classifications among all classifications). Based on k-means clustering, we have identified three clusters of customer type which indicate characteristics of customers who churn. Based on these analyses, our company must monitor the customers who possess/portray these characteristics: (1) high customer service calls with low number of calling minutes, (2) high number of calling minutes, and (3) high number of voicemail messages. This insight is critical to the success of the company, because retaining existing customers has a higher profit return and lower cost with respect to acquiring new customers. As such, our company would benefit in monitoring the customers with the aforementioned characteristics.

**References**

[1] Mechanic, D., & Rochefort, D. A. (1996). Comparative medical systems. Annual Review of Sociology, 22(1), 242.

[2] Young, G. J., Meterko, M., & Desai, K. R. (2000). Patient satisfaction with hospital care: effects of demographic and institutional characteristics. Medical care, 325-334.

[3] Data.Medicare.gov. (2018, July 25). Hospital Compare datasets. Retrieved from <https://data.medicare.gov/data/hospital-compare/>

[4] DeLancey, J. O., Softcheck, J., Chung, J. W., Barnard, C., Dahlke, A. R., & Bilimoria, K. Y. (2017). Associations Between Hospital Characteristics, Measure Reporting, and the Centers for Medicare & Medicaid Services Overall Hospital Quality Star Ratings. Jama, 317(19), 2015-2017.

[5] Tsai, T. C., Jha, A. K., Gawande, A. A., Huckman, R. S., Bloom, N., & Sadun, R. (2015). Hospital board and management practices are strongly related to hospital performance on clinical quality metrics. *Health Affairs*, *34*(8), 1304-1311.

[6] Bloom, N., Propper, C., Seiler, S., & Van Reenen, J. (2015). The impact of competition on management quality: evidence from public hospitals. *The Review of Economic Studies*, *82*(2), 457-489.

[7] Jones, S. S., Rudin, R. S., Perry, T., & Shekelle, P. G. (2014). Health information technology: an updated systematic review with a focus on meaningful use. *Annals of internal medicine*, *160*(1), 48-54.

[8] Kaggle (2018, September 25). Hospital Compare datasets. Retrieved from https://www.kaggle.com/center-for-medicare-and-medicaid/hospital-ratings

[9] Escofier, B., & Pages, J. (1994). Multiple factor analysis (AFMULT package). *Computational statistics & data analysis*, *18*(1), 121-140.

[10] Nadeau, C., & Bengio, Y. (2000). Inference for the generalization error. In *Advances in neural information processing systems* (pp. 307-313).

[11] Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, *5*(Sep), 1089-1105.

[12] Kearns, M., & Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation*, *11*(6), 1427-1453.

[13] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, 111-147.

# Appendix

## Figure 1.1: Summary Statistics of Relevant Numerical Variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attributes** | **Max** | **Min** | **Mean** | **Standard deviation** |
| Hospital Overall Rating (h\_rating) | 5.000 | 1.000 | 2.994 | 0.866 |

## Figure 1.2: Kruskal-Wallis Rank Sum Test

|  |  |  |  |
| --- | --- | --- | --- |
| Attributes (h\_rating by) | Kruskal-Wallis chi-squared | Degrees of Freedom | P-value |
| h\_mortality | 136.15 | 2 | < 2.2e-16 |
| h\_soc | 453.12 | 2 | < 2.2e-16 |
| h\_ra | 673.62 | 2 | < 2.2e-16 |
| h\_pex | 719.14 | 2 | < 2.2e-16 |
| h\_eoc | 79.619 | 2 | < 2.2e-16 |
| h\_toc | 173.25 | 2 | < 2.2e-16 |
| h\_imaging | 11.624 | 2 | 0.002992 |
| h\_ownership | 69.485 | 8 | 6.222e-12 |
| h\_es | 0.053771 | 1 | 0.8166 |

## Figure 1.3: Multiple Factor Analysis (MFA)

## Figure 2.1.1: Summary of key metrics, default dataset

|  |  |  |
| --- | --- | --- |
| **Size of Tree** | **Recall(T)** | **Accuracy** |
| 16 | 66.80% | 84.86% |
| 16 | 65.20% | 84.45% |
| 88 | 71.40% | 87.50% |

\*minNumObj = 50, 20, 2, respectively

## 

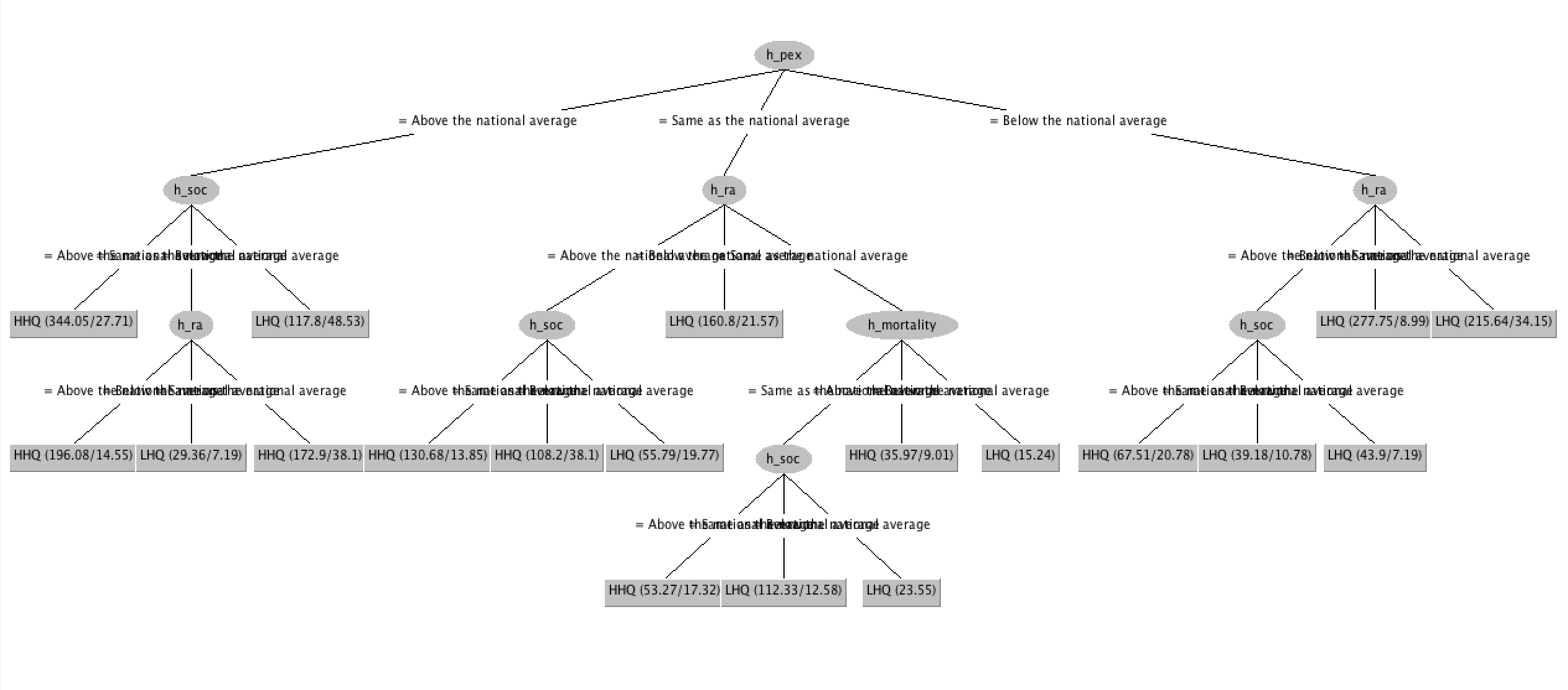
## Figure 2.1.2: Summary of key metrics, balanced dataset

|  |  |  |
| --- | --- | --- |
| **Size of Tree** | **Recall(T)** | **Accuracy** |
| 25 | 80.90% | 82.16% |
| 28 | 83.20% | 82.64% |
| 40 | 87.70% | 83.92% |
| 70 | 90.20% | 85.94% |
| 85 | 89.10% | 85.93% |
| 106 | 89.10% | 86.21% |

\*minNumObj = 40, 30, 20, 10, 5, 2, respectively

## Figure 2.1.3: Key Metrics with regards to the Size of Tree:

## Figure 2.1.4: Final decision tree



## 

## Figure 2.3.1: Model Summary

Call:

glm(formula = Churn. ~ ., family = binomial(link = "logit"),

data = training)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.1806 -0.4982 -0.3264 -0.1821 3.0374

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.1751319 0.8780202 -9.311 < 2e-16 \*\*\*

Account.Length 0.0012835 0.0017095 0.751 0.452743

Int.l.Planyes 2.1459294 0.1794394 11.959 < 2e-16 \*\*\*

VMail.Planyes -1.9383759 0.6873102 -2.820 0.004799 \*\*

VMail.Message 0.0299074 0.0220842 1.354 0.175659

Day.Mins 0.0151153 0.0013500 11.196 < 2e-16 \*\*\*

Day.Calls 0.0022172 0.0033831 0.655 0.512219

Eve.Mins 0.0071118 0.0014274 4.982 6.28e-07 \*\*\*

Eve.Calls 0.0003635 0.0034265 0.106 0.915505

Night.Mins 0.0034712 0.0013641 2.545 0.010938 \*

Night.Calls -0.0058292 0.0035476 -1.643 0.100356

Intl.Mins 0.0917570 0.0250044 3.670 0.000243 \*\*\*

Intl.Calls -0.1187323 0.0322100 -3.686 0.000228 \*\*\*

CustServ.Calls 0.5038960 0.0532189 9.468 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1867.4 on 2302 degrees of freedom

Residual deviance: 1425.6 on 2289 degrees of freedom

AIC: 1453.6

Number of Fisher Scoring iterations: 6

## 

## Figure 2.3.2: Analysis of Deviance Table

Model: binomial, link: logit

Response: Churn.

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL 2302 1867.4

Account.Length 1 0.395 2301 1867.0 0.5298190

Int.l.Plan 1 128.485 2300 1738.5 < 2.2e-16 \*\*\*

VMail.Plan 1 33.488 2299 1705.0 7.172e-09 \*\*\*

VMail.Message 1 1.301 2298 1703.7 0.2539644

Day.Mins 1 130.707 2297 1573.0 < 2.2e-16 \*\*\*

Day.Calls 1 0.359 2296 1572.6 0.5492140

Eve.Mins 1 18.681 2295 1554.0 1.545e-05 \*\*\*

Eve.Calls 1 0.002 2294 1554.0 0.9686894

Night.Mins 1 5.482 2293 1548.5 0.0192164 \*

Night.Calls 1 2.222 2292 1546.2 0.1360864

Intl.Mins 1 11.874 2291 1534.4 0.0005692 \*\*\*

Intl.Calls 1 14.783 2290 1519.6 0.0001206 \*\*\*

CustServ.Calls 1 94.003 2289 1425.6 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

## 

## Figure 2.3.3 Final Model Summary

Call:

glm(formula = Churn. ~ ., family = binomial(link = "logit"),

data = training1)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.0982 -0.5094 -0.3185 -0.1813 3.0965

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.644025 0.634054 -13.633 < 2e-16 \*\*\*

Int.l.Planyes 1.969735 0.176672 11.149 < 2e-16 \*\*\*

VMail.Planyes -1.142282 0.189063 -6.042 1.52e-09 \*\*\*

Day.Mins 0.014999 0.001340 11.193 < 2e-16 \*\*\*

Eve.Mins 0.007695 0.001415 5.438 5.39e-08 \*\*\*

Night.Mins 0.004730 0.001350 3.503 0.000461 \*\*\*

Intl.Mins 0.091873 0.025035 3.670 0.000243 \*\*\*

Intl.Calls -0.116794 0.032231 -3.624 0.000291 \*\*\*

CustServ.Calls 0.471927 0.052752 8.946 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

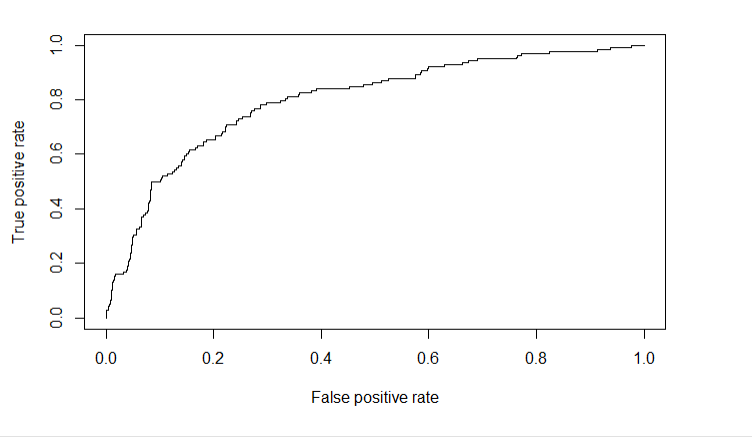
Null deviance: 1867.4 on 2302 degrees of freedom

Residual deviance: 1426.6 on 2294 degrees of freedom

AIC: 1444.6

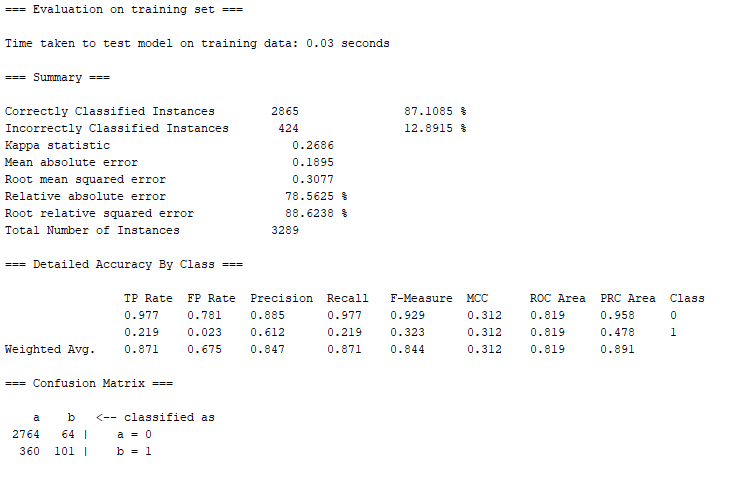
Number of Fisher Scoring iterations: 6

## Figure 2.3.4

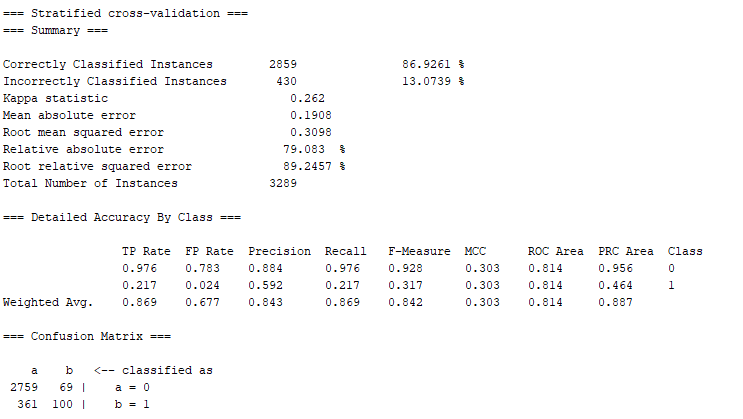


## Figure 2.3.5

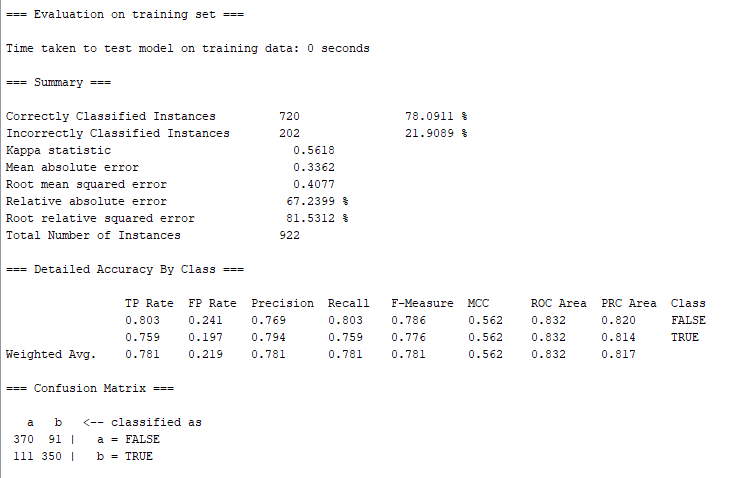
* Using 70% training data (full dataset):



* Use 10-fold Cross-Validation (full dataset):



* Use 70% training data (down-sampling dataset)



## Figure 3.1: Summary of key metrics, default dataset before handling numeric attributes

Note: CV stands for Cross Validation

|  |  |  |
| --- | --- | --- |
| **Test Options** | **Recall(T)** | **Accuracy** |
| Use training set | 31.7% | 87.64% |
| 10-fold CV | 30.4% | 87.40% |
| 3-fold CV | 31.1% | 87.52% |
| 90% Split | 31.4% | 84.99% |

## Figure 3.2: Summary of key metrics, default dataset after handling numeric attributes

Note: KE stands for Kernel Estimator; SD stands for Supervised Discretization; CV stands for Cross Validation

|  |  |  |
| --- | --- | --- |
| **Test Options** | **Recall(T)** | **Accuracy** |
| Use training set (KE) | 17.6% | 87.52% |
| Use training set (SD) | 32.9% | 88.48% |

## 

|  |  |  |
| --- | --- | --- |
| **Test Options** | **Recall(T)** | **Accuracy** |
| 10-fold CV (KE) | 14.5% | 86.77% |
| 10-fold CV (SD) | 26.3% | 87.22% |

|  |  |  |
| --- | --- | --- |
| **Test Options** | **Recall(T)** | **Accuracy** |
| 3-fold CV (KE) | 14.9% | 86.62% |
| 3-fold CV (SD) | 30.4% | 87.19% |

## 

|  |  |  |
| --- | --- | --- |
| **Test Options** | **Recall(T)** | **Accuracy** |
| 90% split (KE) | 15.7% | 86.19% |
| 90% split (SD) | 27.5% | 84.99% |

## 

## Figure 3.3: Summary of key metrics, down-sampled dataset before handling numeric attributes

Note: CV stands for Cross Validation

|  |  |  |
| --- | --- | --- |
| **Test Options** | **Recall(T)** | **Accuracy** |
| Use training set | 82.6% | 81.78% |
| 10-fold CV | 82.2% | 81.02% |
| 3-fold CV | 80.0% | 79.72% |
| 90% Split | 84.4% | 84.78% |

## 

## Figure 3.4.1: Summary of key metrics, down-sampled dataset after handling numeric attributes

|  |  |  |
| --- | --- | --- |
| **Test Options** | **Recall(T)** | **Accuracy** |
| Use training set (KE) | 82.4% | 84.60% |
| Use training set (SD) | 76.4% | 81.24% |

## 

|  |  |  |
| --- | --- | --- |
| **Test Options** | **Recall(T)** | **Accuracy** |
| 10-fold CV (KE) | 78.5% | 80.80% |
| 10-fold CV (SD) | 74.8% | 80.37% |

|  |  |  |
| --- | --- | --- |
| **Test Options** | **Recall(T)** | **Accuracy** |
| 3-fold CV (KE) | 77.7% | 79.72% |
| 3-fold CV (SD) | 76.6% | 80.04% |

## 

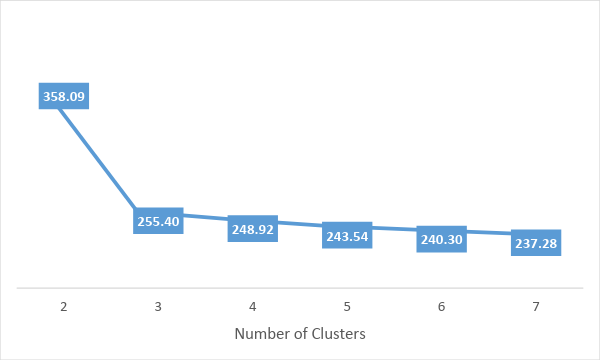
|  |  |  |
| --- | --- | --- |
| **Test Options** | **Recall(T)** | **Accuracy** |
| 90% split (KE) | 71.1% | 80.43% |
| 90% split (SD) | 84.4% | 85.87% |

## 

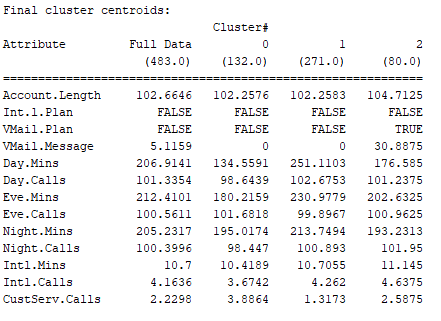
## Figure 3.4.2: Summary of key metrics, down-sampled dataset after handling numeric attributes

|  |  |  |
| --- | --- | --- |
| **Test Options** | **Recall(T)** | **Accuracy** |
| 90% split (SD) | 84.4% | 85.87% |
| Use training set (KE) | 82.4% | 84.60% |
| 10-fold CV | 82.2% | 81.02% |
| 3-fold CV (SD) | 76.6% | 80.04% |

## Figure 4.1.1:



## Figure 4.1.2:



## Figure 4.1.3:

