Project Status Report

Business Understanding

Telecommunication industry in the past couple of years have faced pricing pressure, with the disappearance of the differentiability of product and service between network has led customer to lose loyalty and become indifferent between telecommunication companies. This combined with the increase competition from cable operators and startups, has forced many telecommunication companies to attempt to retain their existing customer with more competitive offers, bundles and price cuts.

Our client, an incumbent in the telecommunication industry, wish to be able to identify and predict if a customer will churn. As the on average the cost to acquire a new customer is much more expensive to retain a customer, our client thus wishes to use data science models to build a model that is able to identify potential churning customer and target them those customers with a new competitive offers/bundles or other marketing campaign in an attempt to retain them.

Data Understanding & Visualization

Taking a cross sectional subset of the database of customer information on a particular month from our client’s database, with each row of dataset to represent a customer and each column containing customer attributes of the following:

* Services that each customer has signed up for
* Customer account information
* Demographic information about customer
* Target variable: whether customer has left within the last month

In total there are 7043 observation of 20 variables, with 16 categorical variables and 3 continuous variable and 1 target variable:

1. 16 Categorical variables:
   1. 6 binary variables (Gender, Senior Citizen, Partner, Dependent, Phone Service, and Paperless billing)
   2. 9 3 -Factor level variables (Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming movies and Contact)
   3. 1 4-Factor level variable (Payment method)
2. 3 Continuous variables:
   1. Tenure, Monthly Charges, Total Charges
3. 1 Target Variable
   1. Churn

Missing data Exploration:

From Figure 1., we can observe that out of 7043 observation, there are only 11 missing values and all of them belong to missing total charges column. Taking a further dive in these 11 rows (table 1), we are able to observe a commonality between these 11 customer is that tenure = 0, meaning that maybe this may be their first month with the company and thus our client hasn’t billed them.

Data Visualization:

Figure 2. Churn Percentage: around 26% of customer left our client within the past month

Taking a further dive at our categorical variables with respect to churn, we can observe the following trend:

* Senior Citizens churn percentage are higher (Figure 2)
* Customer with dependents or partners tend to have lower churn rate compared to counterparts (Figure 3)
* Customer with paperless billing have higher churn rate (Figure 3)
* Customer with Fiber Optic Internet Service have significant higher churn rate
* Customer with No online security, or online backup or tech support have higher churn rate (Figure 4)
* Customer with monthly subscription are more likely to churn compared to customer with one- or two-year contract (Figure 4)
* Customer with Electronic Check payment method tend to leave our client more compared to other options. (Figure 5)

Taking a further dive at our continuous variable with respect to churn (Figure 6 and 7), we can observe the following trend:

* Recent client is more likely to churn
* Clients with higher monthly charges are more likely to churn
* There doesn’t seem to many outliers

Modelling

Models used in analysis:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Linear SVM

Threshold/Cut off Point Analysis:

As we are attempting to identify customer that are going to churn, we thus need to focus on sensitivity metric compared to accuracy. As it is comparatively more expensive to acquire customer than retain customer, thus we are not as concern with false positive, but rather concerned with false negative. We would ideally like a model that is able to successful target all customer that are going to churn, and it should matter less if we have a higher number of false positive to us a telecommunication company. Thus, we should have a lower threshold value than 0.5, though the actual value often requires domain knowledge which we lack, thus we are going to use a more objective method to set out threshold value.

From Figure 8, we are going to use the intercept point of accuracy, sensitivity and specificity as our cutoff point for the evaluation of our models. (Cutoff Value = 0.2892929)

Choice of model and Pros and Cons:

Results Interpretation:

Logistic Regression (post model selection and VIF):

* Refer to table 2: GLM Summary Output

Decision Tree Model:

* Refer to Figure 8

Random Forest Model:

* Refer to Table 3

Linear SVM Model:

* Refer to Table 4

Evaluation

In general, we can access the performance of our models using AUC and sensitivity metric.

* Define AUC and Sensitivity
* Refer to figure 10 and table 5

Confusion matrix for all of them

* Refer to table 6 ~ table 9
* Problem faced – unable to change cutoff value for SVM model -> comes out as yes no and not probability?

Business Case

* CLV and ROMI

Deployment

Developing a model to predict potential customer churn is just the first step, next telecommunication company would have to design a marketing campaign/promotion and target those users.

Risk analysis

Appendix A: Figures

Figure 1: Missing Data Analysis

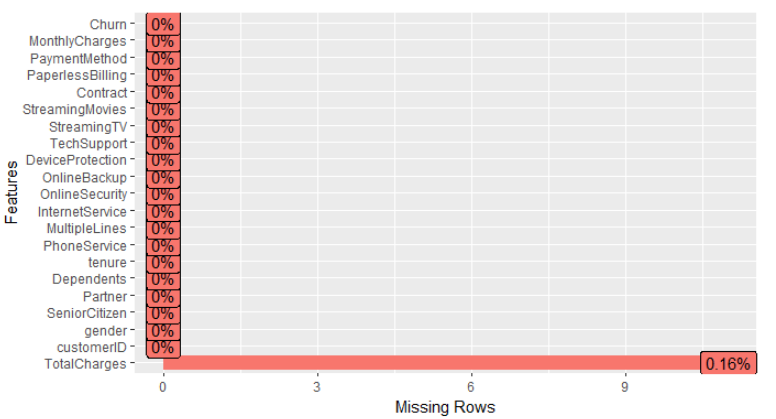


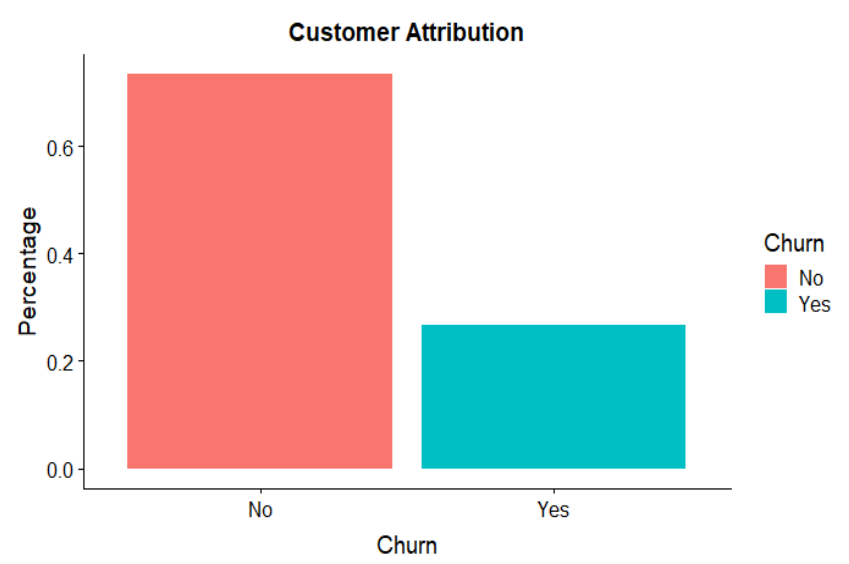
Figure 2: Customer Churn/Attribution percentage

Figure 3: Binary variable analysis

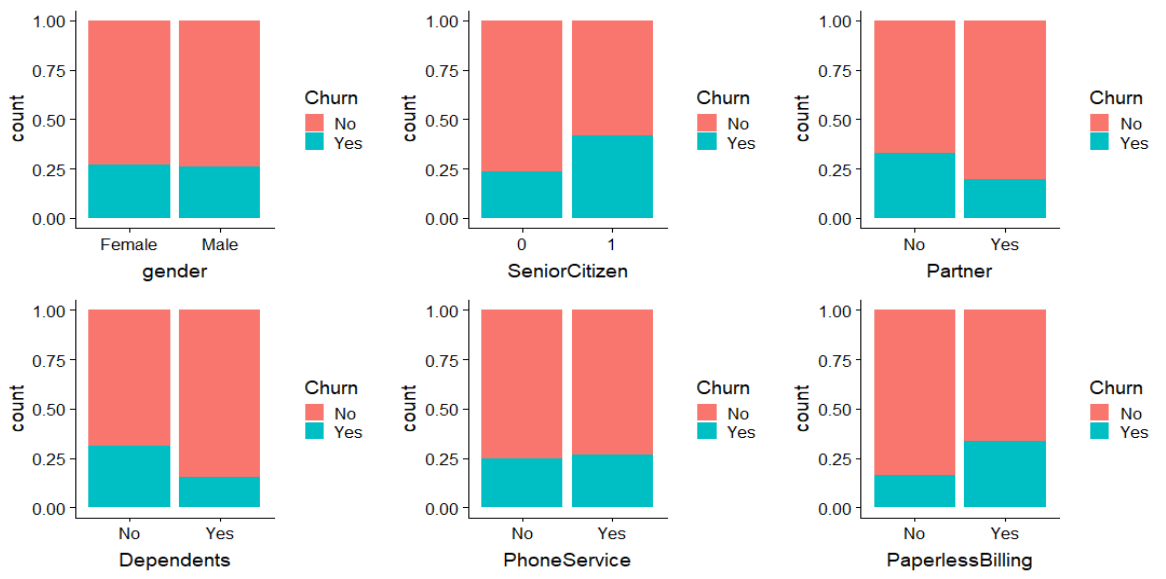


Figure 4: 3-factor variable analysis

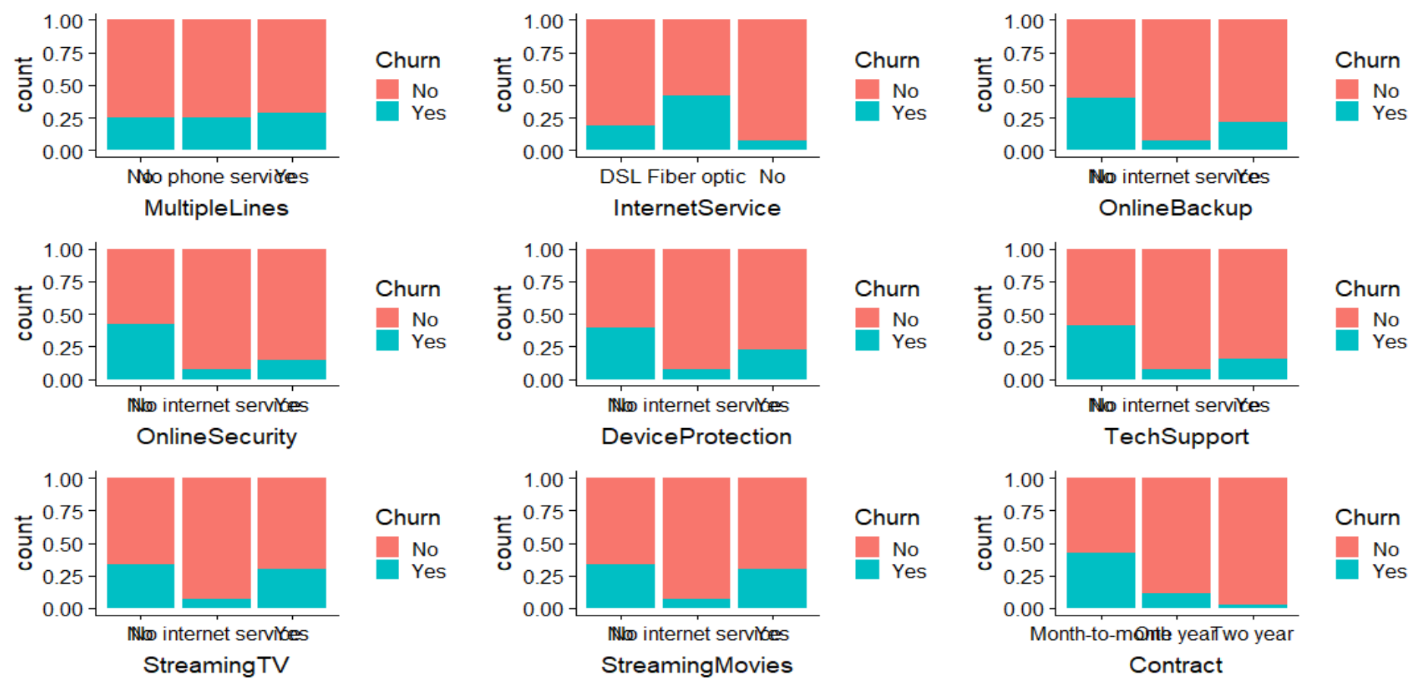


Figure 5: 4-factor variable analysis

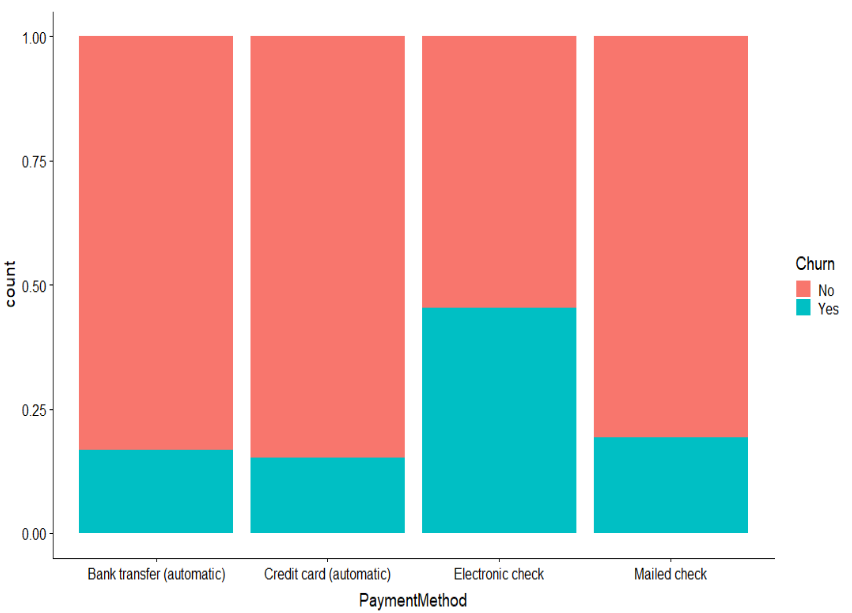


Figure 6: Continuous variable analysis

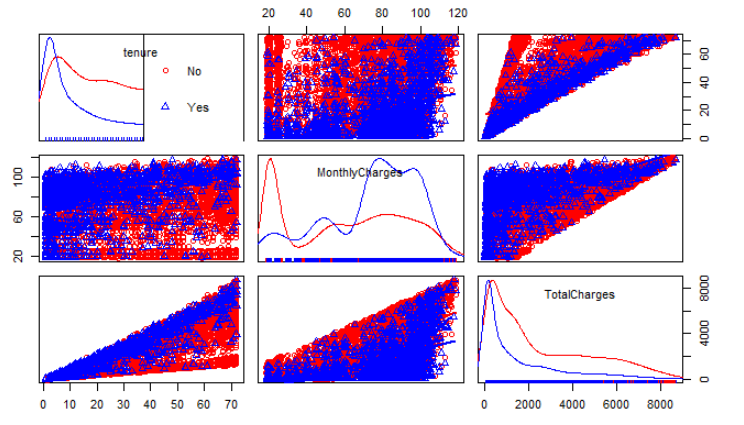


Figure 7: continuous variable analysis

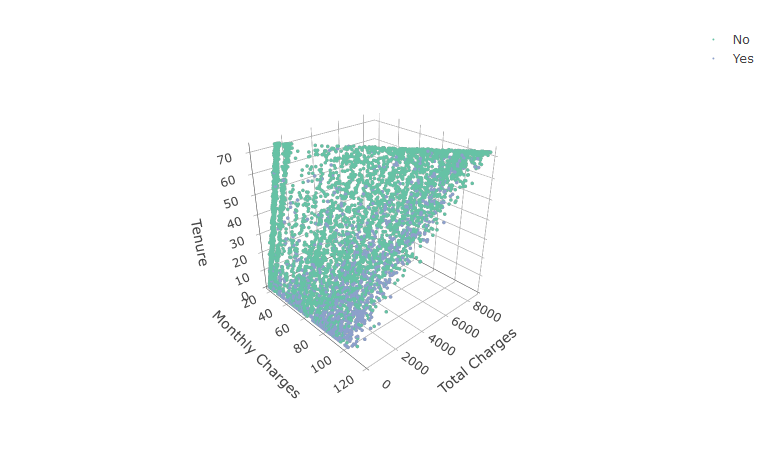


Figure 8: Pruned Decision Tree Model:

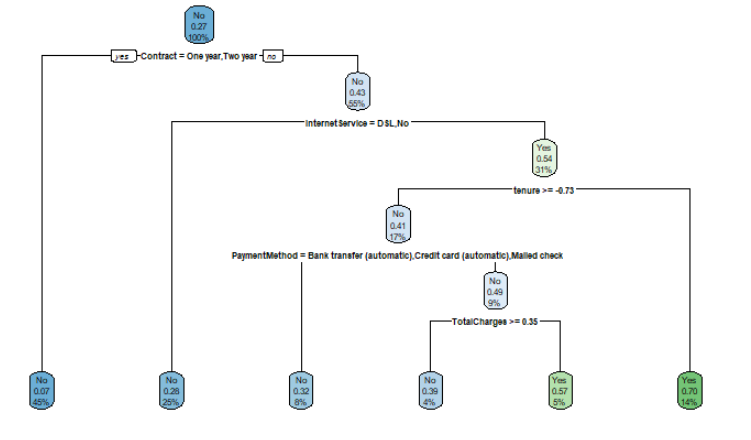


Figure 9: Threshold analysis

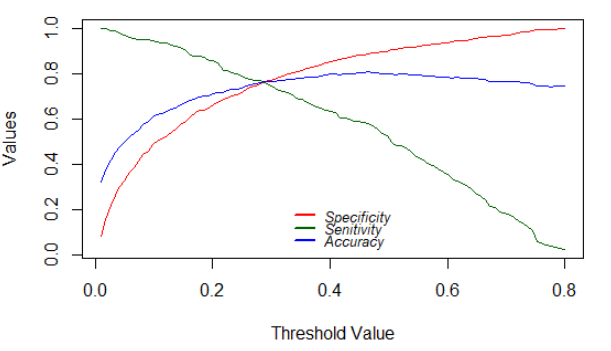
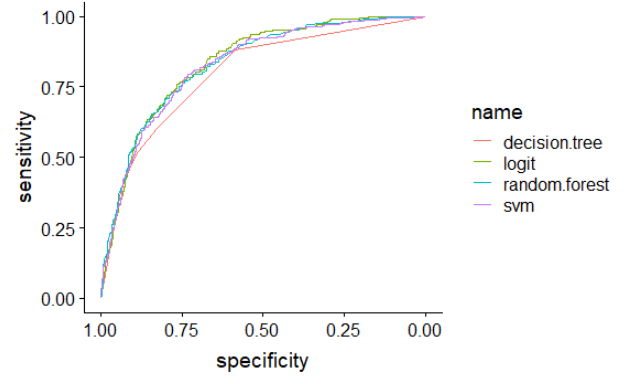


Figure 10: ROC curve



Appendix B: Tables

Table 1: Missing Data Table

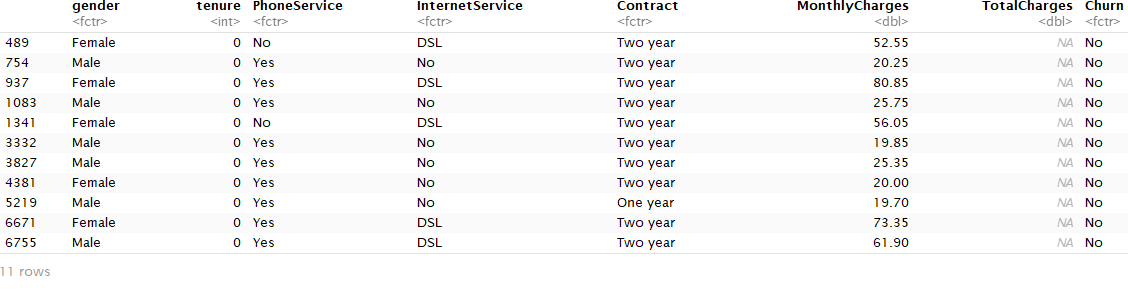


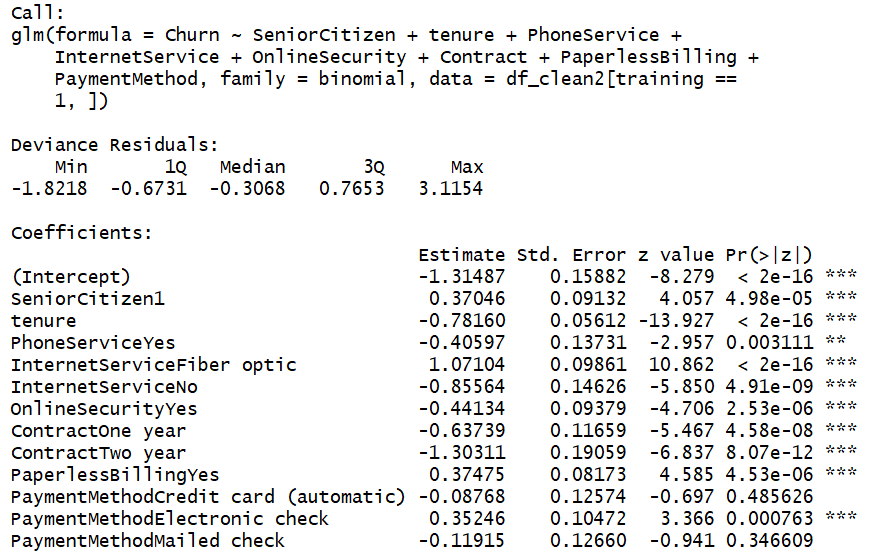
Table 2: Logistic Regression Output

Table 3: Random Forest Output:

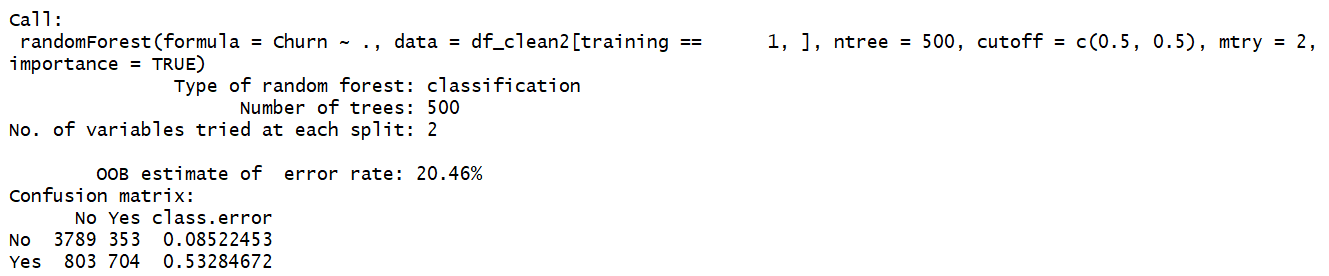


Table 4: Linear SVM Output

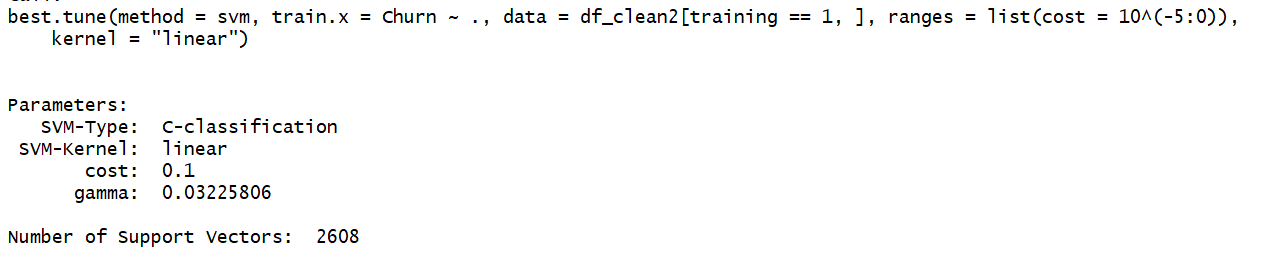


Table 5: AUC Summary Table



Table 6: Logistic Regression Confusion Matrix

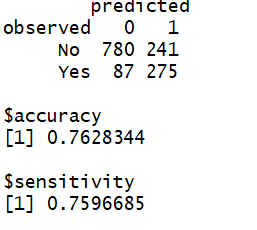


Table 7: Decision Tree Confusion Matrix

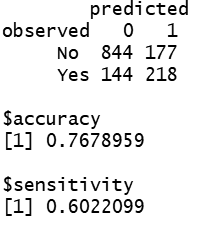


Table 8: Random Forest Confusion Matrix

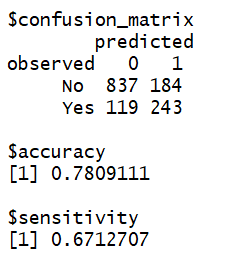


Table 9: