ITM 883

Business Analytics and Problem Solving

Project Report
Customer Churn Analysis

TEAM OUTLIERS:

Karan Dalal	-	dalalkar@msu.edu
Neil Joshi	-	joshine2@msu.edu
Syed Kashif Mujtaba	-	kamoonpu@msu.edu
Vishal Agarwal	-	agarwa97@msu.edu

1. OBJECTIVE

In this project, we will analyze the historical customer data of a Telco organization to identify customers who are likely to churn and the factors that lead to a customer churning. We will be using logistic regression to classify whether a customer will churn or not.

2. DATA OVERVIEW AND DESCRIPTION

The dataset for this project has been obtained from Kaggle.com. The dataset has over 7000 rows and 20 predictor variables. Each row represents a customer and each column contains customer's attributes.

COLUMN NAME	COLUMN DESCRIPTION
Customer ID	A unique ID for each customer
Gender	Whether the customer is a male or a female
SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)
Partner	Whether the customer has a partner or not (Yes, No)
Dependents	Whether the customer has dependents or not (Yes, No)
Tenure	Number of months the customer has stayed with the company
PhoneService	Whether the customer has a phone service or not (Yes, No)
MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)
InternetService	Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract	The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	The customer's payment method (Electronic check, mailed check, Bank transfer, Credit card)
MonthlyCharges	The amount charged to the customer monthly
TotalCharges	The total amount charged to the customer (Combination of tenure and monthly charges)
Churn	Whether the customer churned or not (Yes or No)

3. METHODOLOGY

i. Exploratory Data Analysis

In this phase, we performed exploratory data analysis to get insights using ggplot2 library.

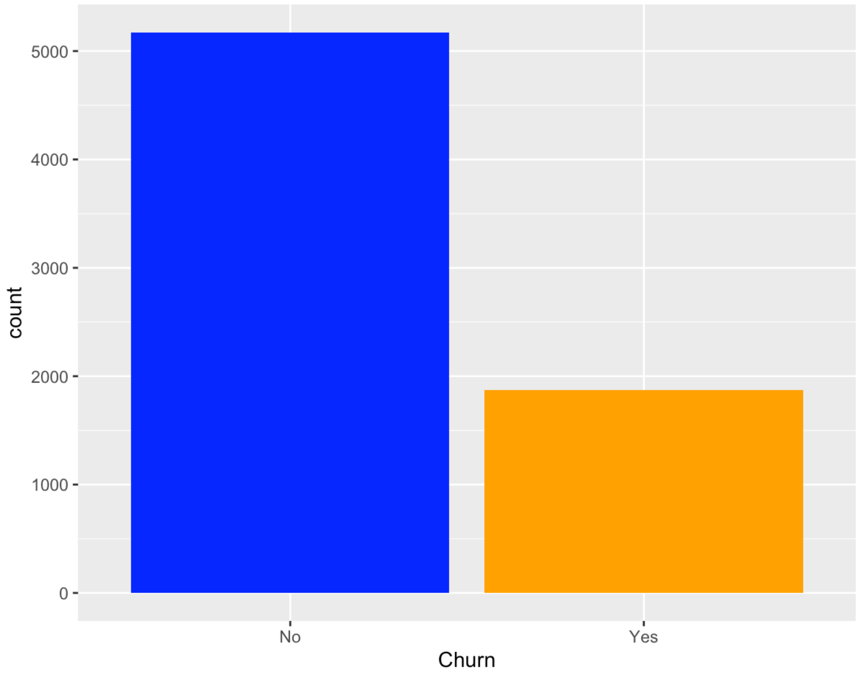


Fig. 1: The organization is successful in retaining majority of its customers

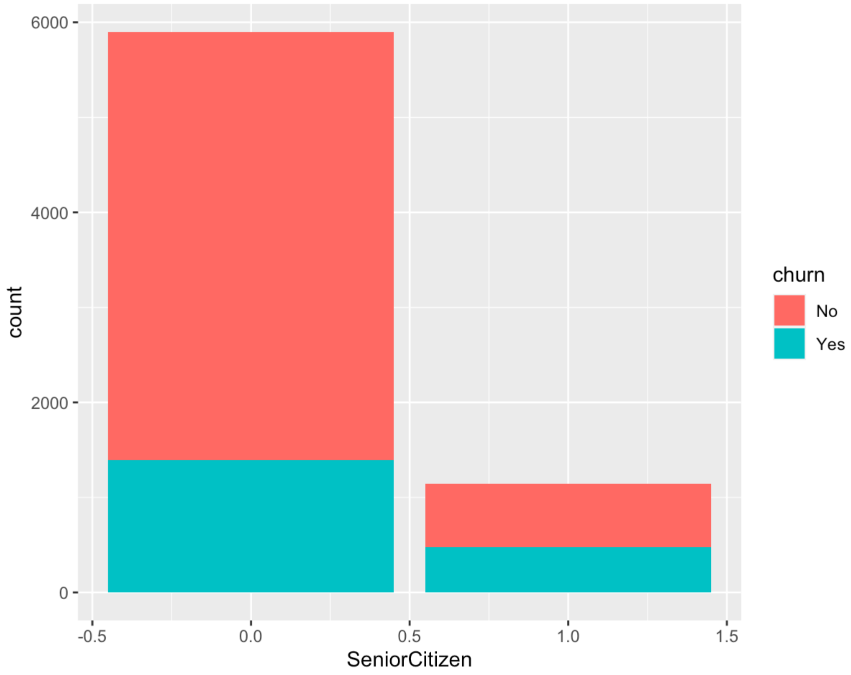


Fig. 2: Majority of the customers that churned were not senior citizens

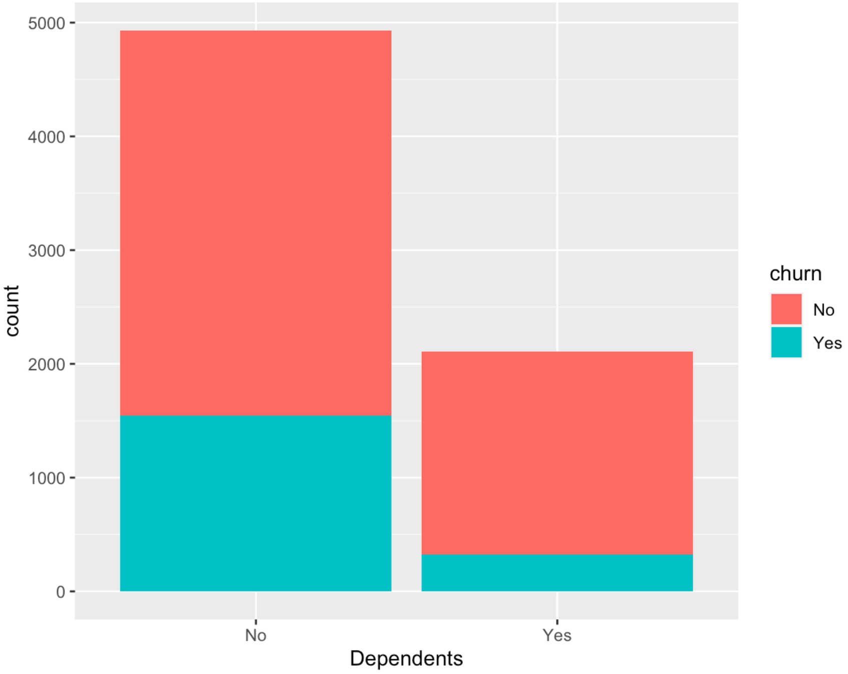


Fig. 3: Majority of the customers that churned did not have dependents

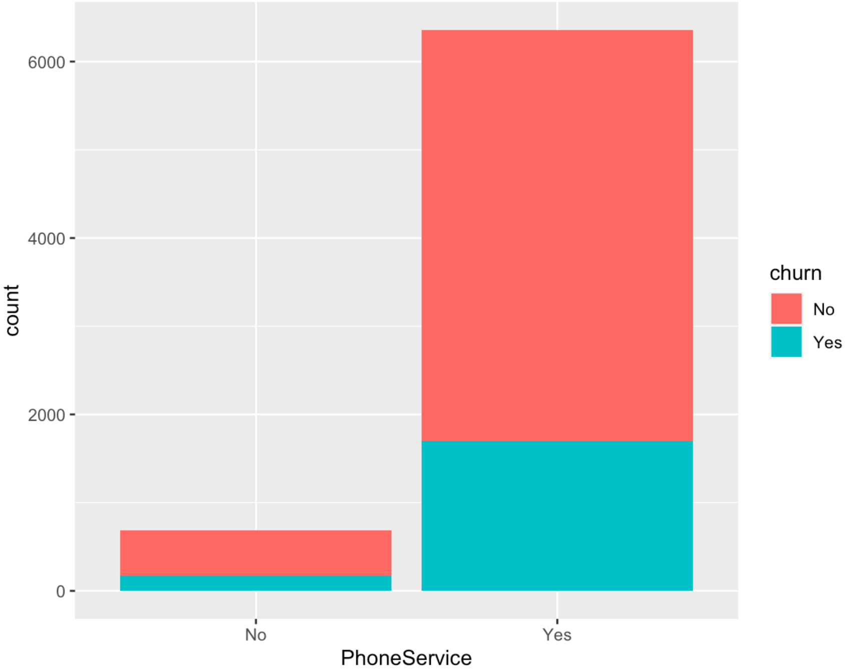


Fig. 4: Majority of the customers that churned had phone service

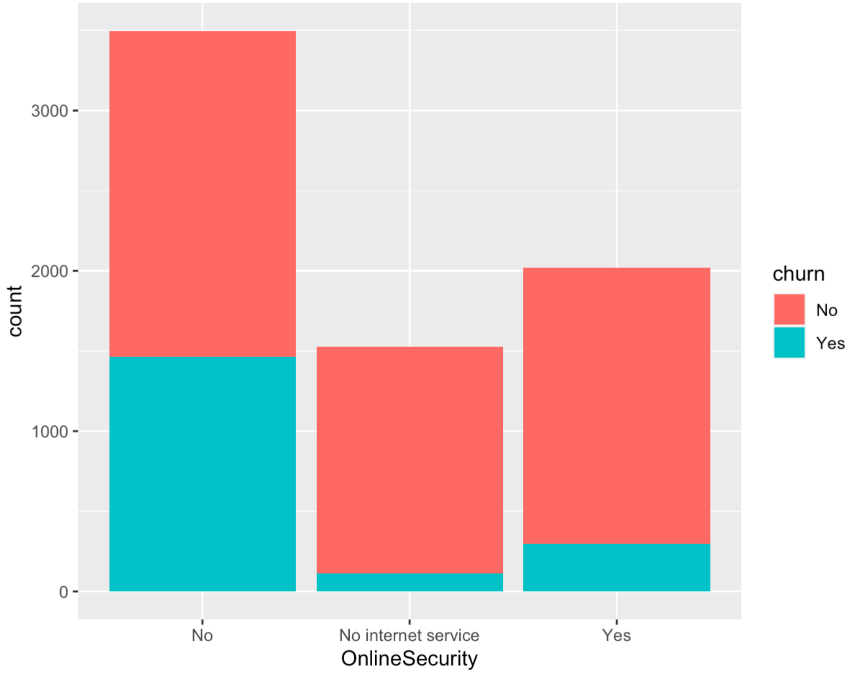


Fig. 5: Majority of the customers that churned did not have Online Security

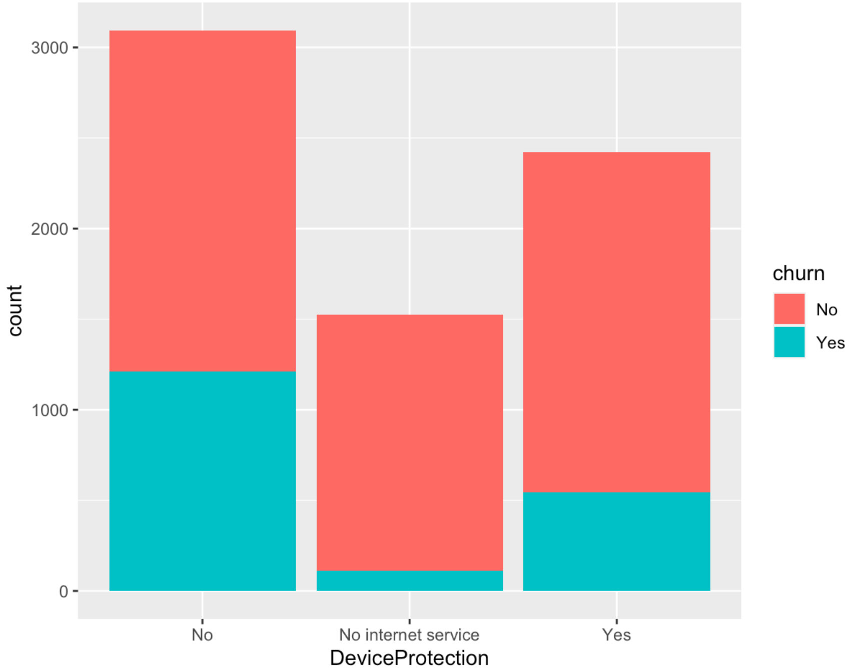


Fig. 6: Majority of the customers that churned did not have device protection

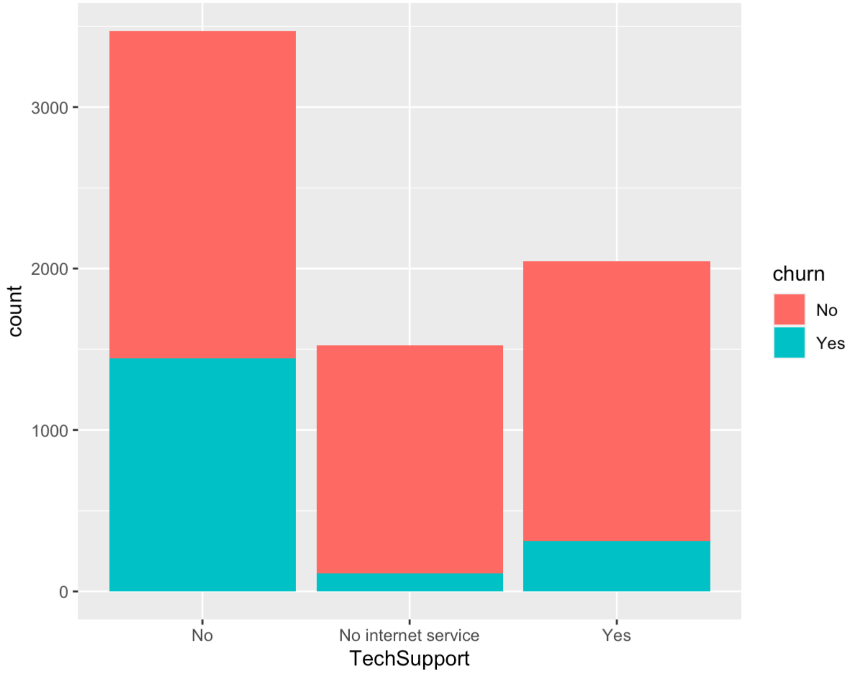


Fig. 7: Majority of the customers that churned did not have tech support

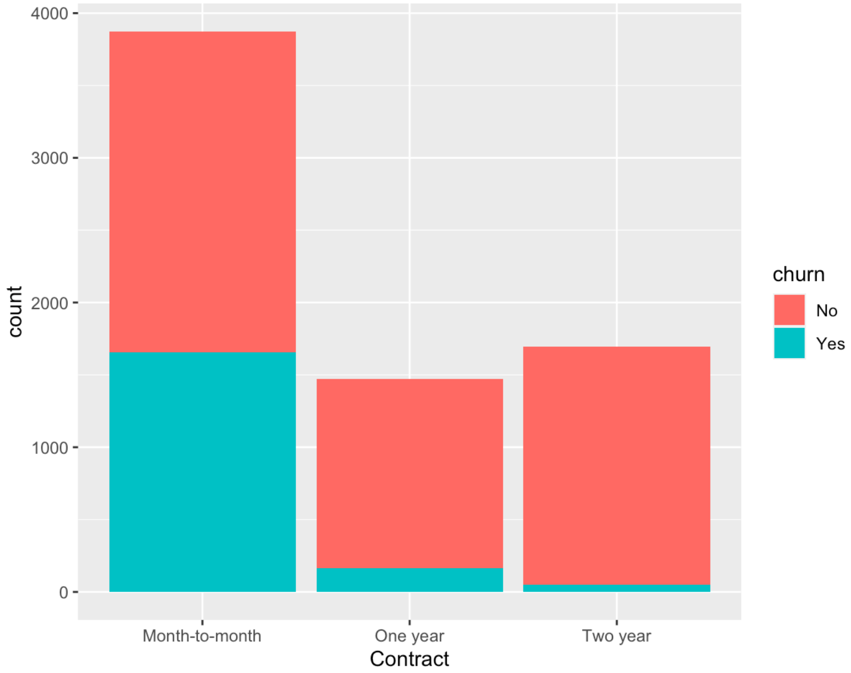


Fig. 8: Majority of the customers that churned had month-to-month contract

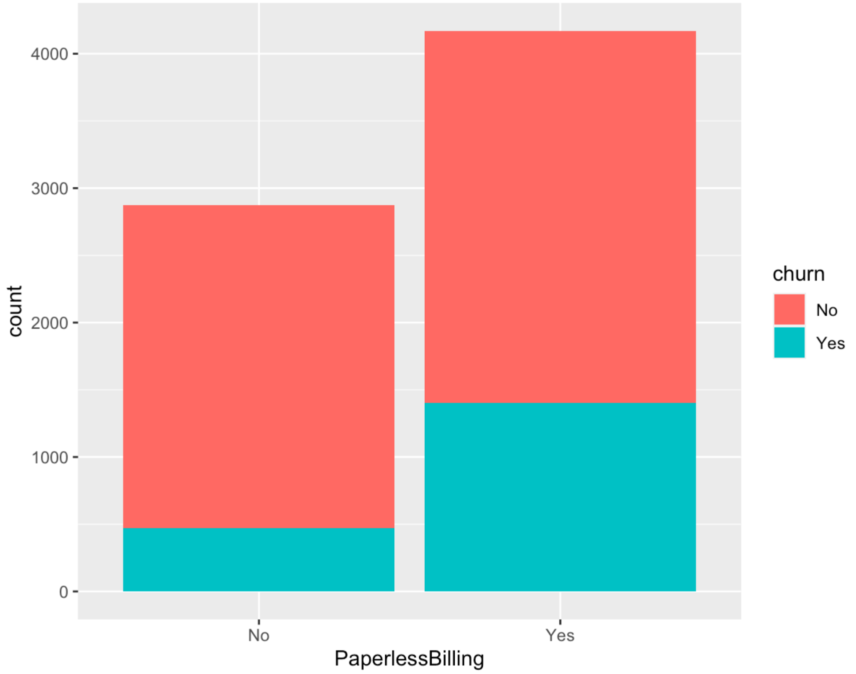


Fig. 9: Majority of the customers that churned had paperless billing

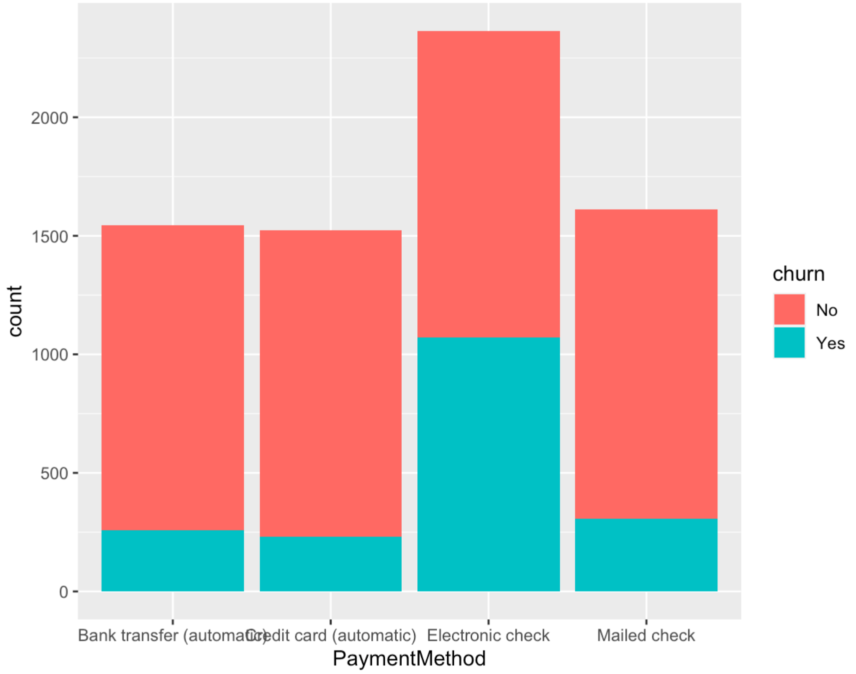


Fig. 10: Majority of the customers that churned paid through electronic check

ii. Data Preprocessing

We performed the following tasks during this phase of the project:

- Identified and imputed missing values
- Converted dependent variable (Churn) to binary format as follows:
  - 0: The customer did not churn, i.e. he/she is still with the company
  - 1: The customer churned, i.e. he/she is no longer a customer
- Converted numeric categorical variables into factors using as.factor()
- Removed multi-collinearity from the dataset

iii. Creating Logistic Regression Models

For our analysis we created the following models:

- <u>Model 1:</u> Logistic Regression model using all the independent variable
- <u>Model 2:</u> Reduced Logistic Regression model using stepAIC()
- <u>Model 3:</u> Logistic Regression model with features selected using exhaustive search

For each of these models, the threshold value for classification has been selected using the ROC curve.

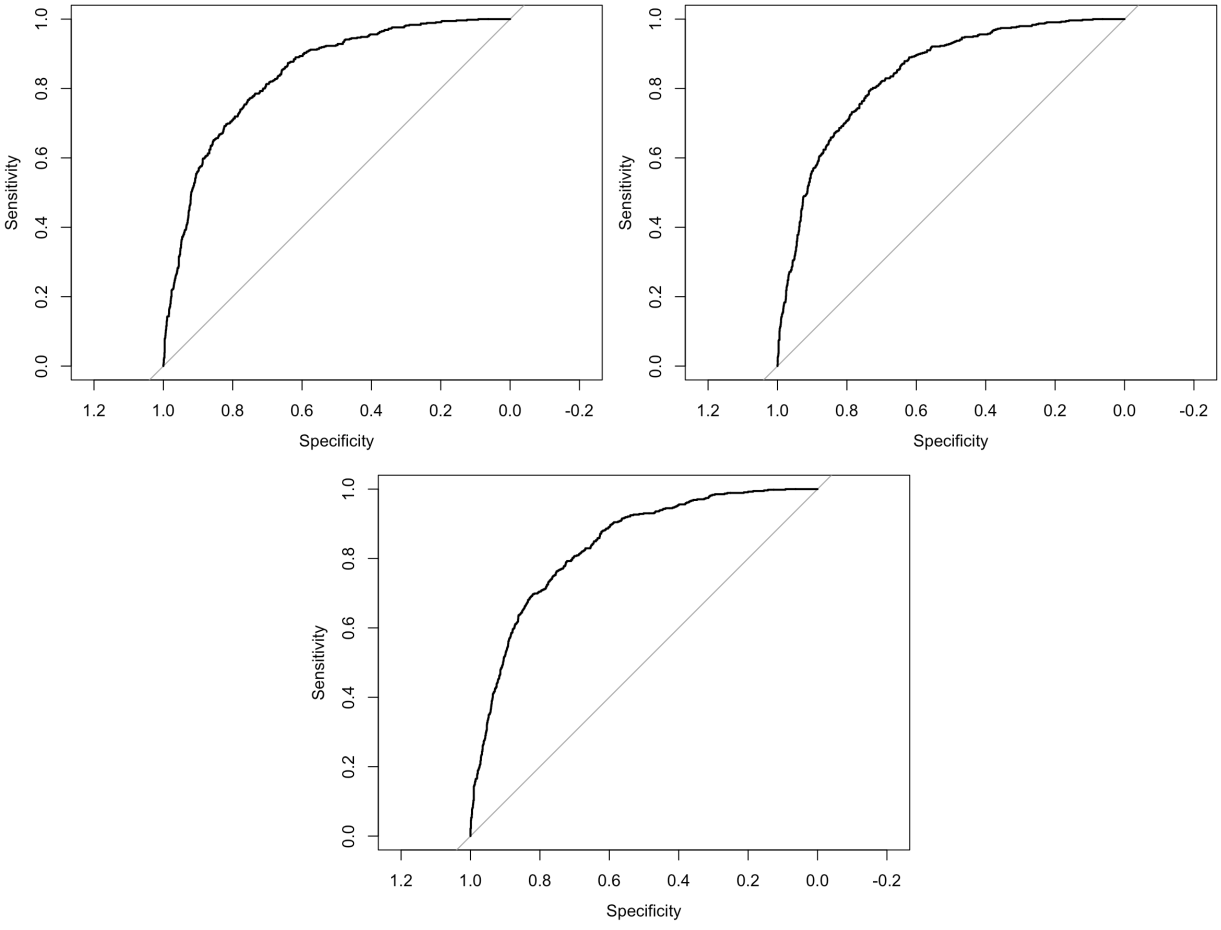
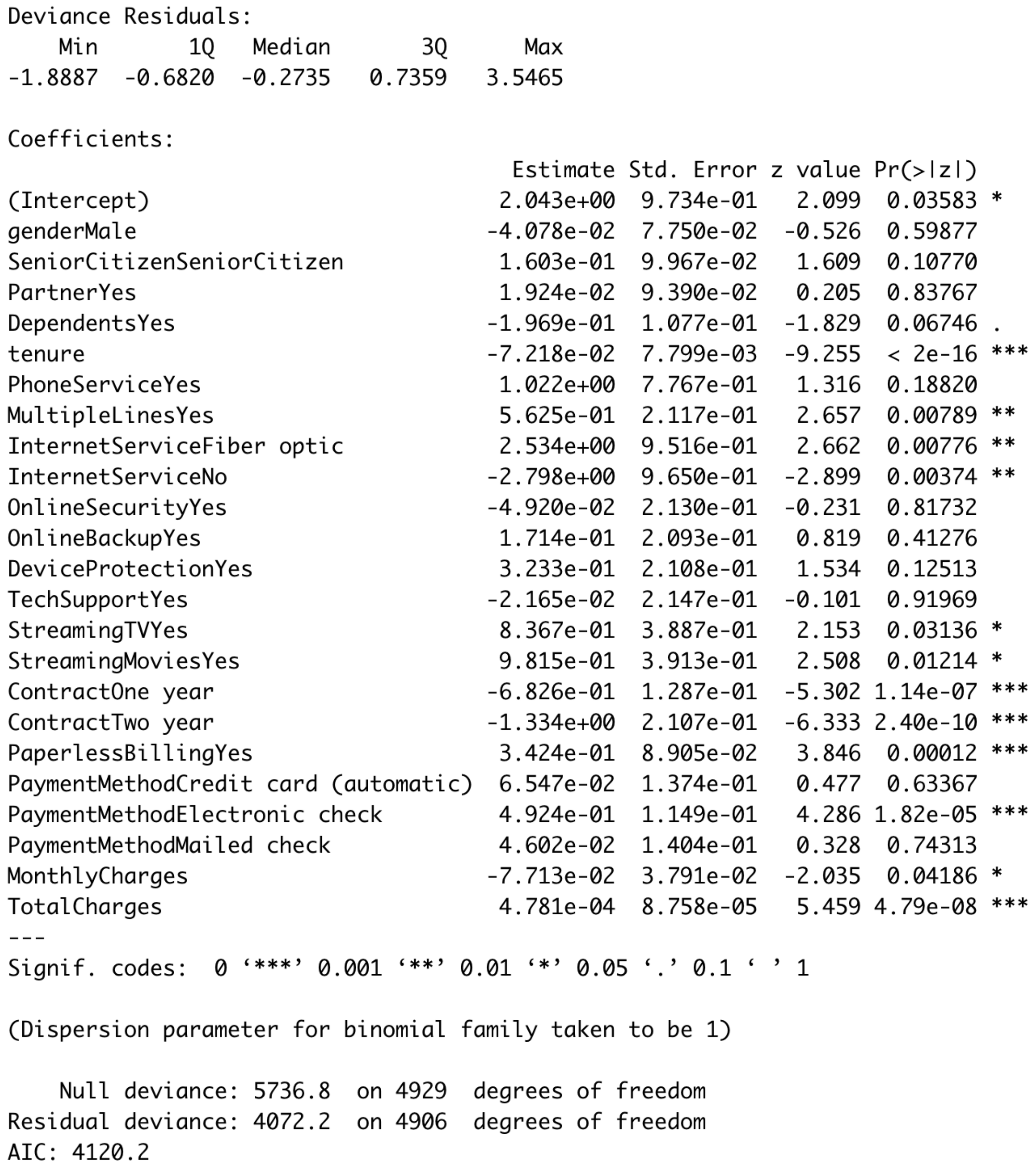


Fig. 11: ROC curve for Model 1, Model 2 and Model 3 respectively

Model 1:

This model uses 23 independent variables to predict if a customer will churn or not. The best threshold value for this model is 0.286. The summary of the model is as follows:

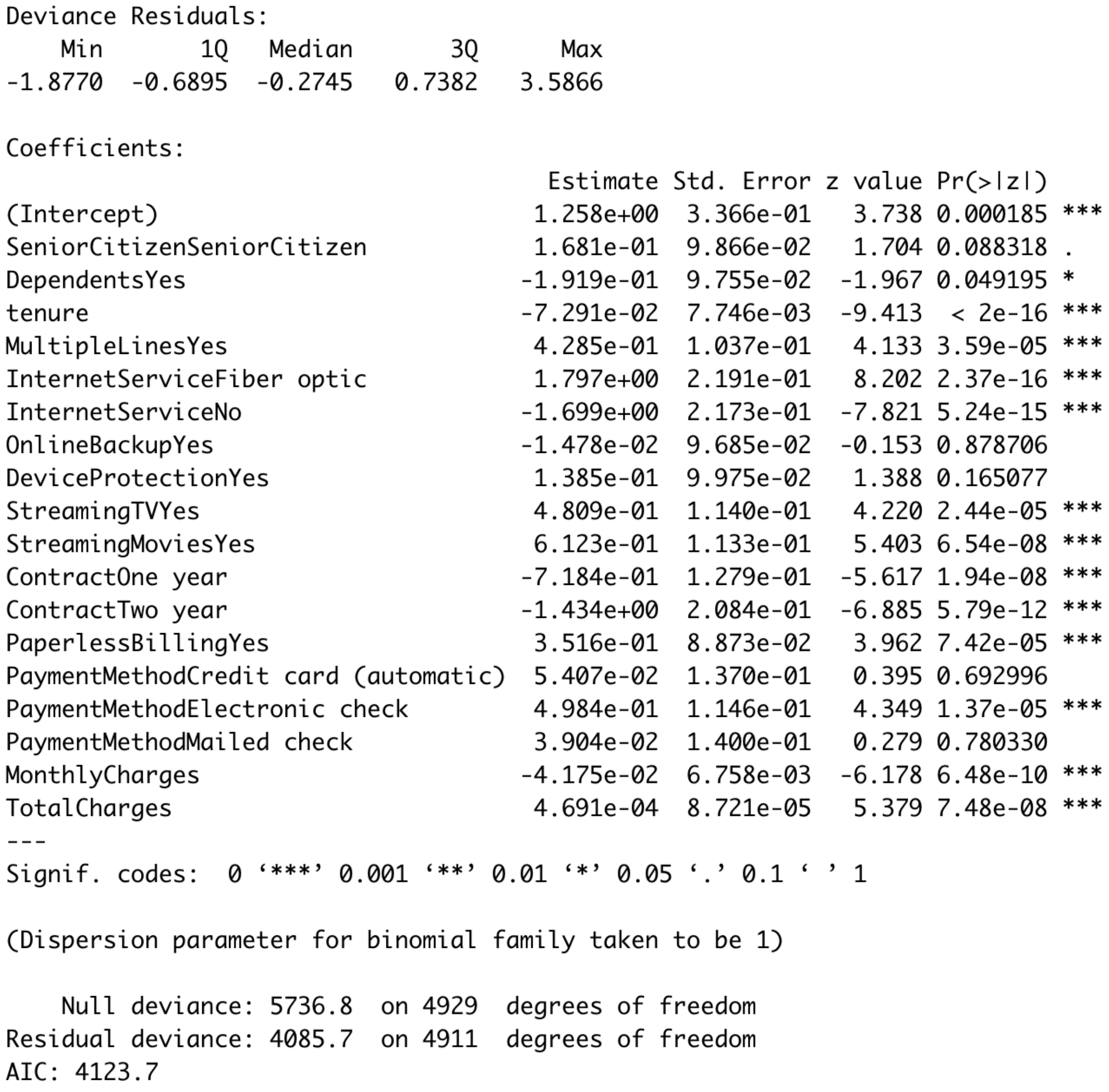


```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8887  -0.6820  -0.2735   0.7359   3.5465

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                           2.043e+00  9.734e-01   2.099  0.03583 *
genderMale                           -4.078e-02  7.750e-02  -0.526  0.59877
SeniorCitizenSeniorCitizen            1.603e-01  9.967e-02   1.609  0.10770
PartnerYes                            1.924e-02  9.390e-02   0.205  0.83767
DependentsYes                        -1.969e-01  1.077e-01  -1.829  0.06746 .
tenure                               -7.218e-02  7.799e-03  -9.255  < 2e-16 ***
PhoneServiceYes                       1.022e+00  7.767e-01   1.316  0.18820
MultipleLinesYes                      5.625e-01  2.117e-01   2.657  0.00789 **
InternetServiceFiber optic            2.534e+00  9.516e-01   2.662  0.00776 **
InternetServiceNo                    -2.798e+00  9.650e-01  -2.899  0.00374 **
OnlineSecurityYes                    -4.920e-02  2.130e-01  -0.231  0.81732
OnlineBackupYes                       1.714e-01  2.093e-01   0.819  0.41276
DeviceProtectionYes                   3.233e-01  2.108e-01   1.534  0.12513
TechSupportYes                       -2.165e-02  2.147e-01  -0.101  0.91969
StreamingTVYes                        8.367e-01  3.887e-01   2.153  0.03136 *
StreamingMoviesYes                    9.815e-01  3.913e-01   2.508  0.01214 *
ContractOne year                     -6.826e-01  1.287e-01  -5.302 1.14e-07 ***
ContractTwo year                     -1.334e+00  2.107e-01  -6.333 2.40e-10 ***
PaperlessBillingYes                   3.424e-01  8.905e-02   3.846  0.00012 ***
PaymentMethodCredit card (automatic)  6.547e-02  1.374e-01   0.477  0.63367
PaymentMethodElectronic check         4.924e-01  1.149e-01   4.286 1.82e-05 ***
PaymentMethodMailed check             4.602e-02  1.404e-01   0.328  0.74313
MonthlyCharges                       -7.713e-02  3.791e-02  -2.035  0.04186 *
TotalCharges                          4.781e-04  8.758e-05   5.459 4.79e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5736.8  on 4929  degrees of freedom
Residual deviance: 4072.2  on 4906  degrees of freedom
AIC: 4120.2
```

Fig. 12: Summary of Model 1

Model 2:

This model was created by reducing Model 1 using the stepAIC() function. We have used ‘both’ direction to reduce this model. It has 18 independent variables. The best threshold value for this model is 0.27. The summary of the model is as follows:

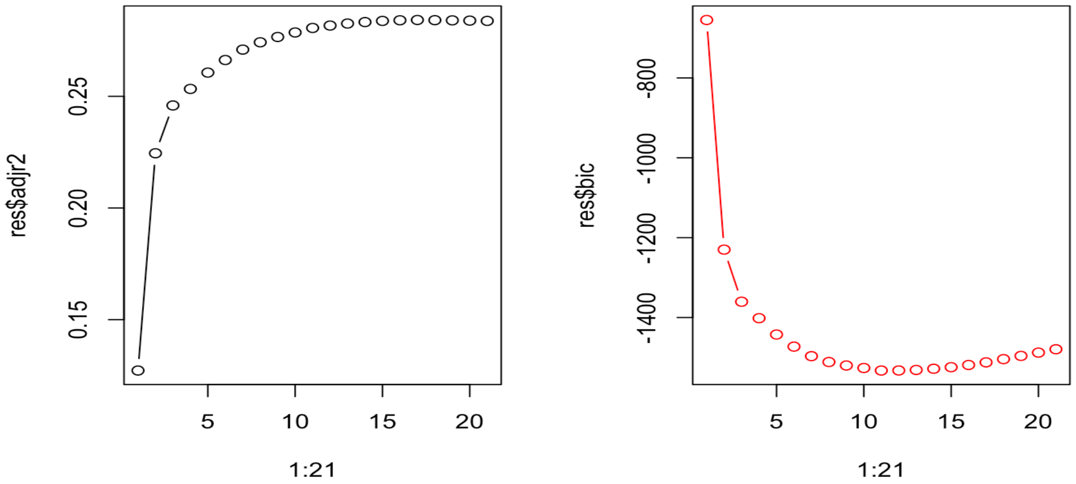


```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8770  -0.6895  -0.2745   0.7382   3.5866

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                           1.258e+00  3.366e-01   3.738 0.000185 ***
SeniorCitizenSeniorCitizen            1.681e-01  9.866e-02   1.704 0.088318 .
DependentsYes                        -1.919e-01  9.755e-02  -1.967 0.049195 *
tenure                               -7.291e-02  7.746e-03  -9.413  < 2e-16 ***
MultipleLinesYes                      4.285e-01  1.037e-01   4.133 3.59e-05 ***
InternetServiceFiber optic            1.797e+00  2.191e-01   8.202 2.37e-16 ***
InternetServiceNo                    -1.699e+00  2.173e-01  -7.821 5.24e-15 ***
OnlineBackupYes                      -1.478e-02  9.685e-02  -0.153 0.878706
DeviceProtectionYes                   1.385e-01  9.975e-02   1.388 0.165077
StreamingTVYes                        4.809e-01  1.140e-01   4.220 2.44e-05 ***
StreamingMoviesYes                    6.123e-01  1.133e-01   5.403 6.54e-08 ***
ContractOne year                     -7.184e-01  1.279e-01  -5.617 1.94e-08 ***
ContractTwo year                     -1.434e+00  2.084e-01  -6.885 5.79e-12 ***
PaperlessBillingYes                   3.516e-01  8.873e-02   3.962 7.42e-05 ***
PaymentMethodCredit card (automatic)  5.407e-02  1.370e-01   0.395 0.692996
PaymentMethodElectronic check         4.984e-01  1.146e-01   4.349 1.37e-05 ***
PaymentMethodMailed check             3.904e-02  1.400e-01   0.279 0.780330
MonthlyCharges                       -4.175e-02  6.758e-03  -6.178 6.48e-10 ***
TotalCharges                          4.691e-04  8.721e-05   5.379 7.48e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5736.8  on 4929  degrees of freedom
Residual deviance: 4085.7  on 4911  degrees of freedom
AIC: 4123.7
```

Fig. 13: Summary of Model 2

Model 3:

This model has been created using the exhaustive search method. From the two graphs below, we can see that the best model should have 12 independent variables. The best threshold value for this model is 0.38. The summary of the model is as follows:

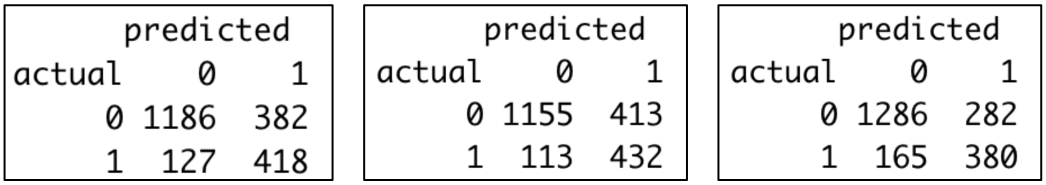


```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7979  -0.6756  -0.2800   0.7453   3.5198

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -4.016e-01  1.523e-01  -2.637  0.00837 **
tenure                               -6.643e-02  7.454e-03  -8.913  < 2e-16 ***
MultipleLinesYes                      1.422e-01  9.290e-02   1.530  0.12593
InternetServiceFiber optic            5.313e-01  1.092e-01   4.867 1.13e-06 ***
InternetServiceNo                    -9.822e-01  1.577e-01  -6.228 4.74e-10 ***
OnlineSecurityYes                    -4.533e-01  1.006e-01  -4.505 6.64e-06 ***
TechSupportYes                       -4.352e-01  1.019e-01  -4.271 1.94e-05 ***
StreamingMoviesYes                    2.776e-01  9.333e-02   2.974  0.00294 **
ContractOne year                     -7.301e-01  1.273e-01  -5.737 9.64e-09 ***
ContractTwo year                     -1.376e+00  2.087e-01  -6.592 4.35e-11 ***
PaperlessBillingYes                   3.777e-01  8.819e-02   4.282 1.85e-05 ***
PaymentMethodCredit card (automatic)  7.333e-02  1.367e-01   0.536  0.59181
PaymentMethodElectronic check         5.257e-01  1.142e-01   4.604 4.13e-06 ***
PaymentMethodMailed check             4.483e-02  1.394e-01   0.322  0.74773
TotalCharges                          4.018e-04  8.173e-05   4.917 8.80e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5736.8  on 4929  degrees of freedom
Residual deviance: 4101.4  on 4915  degrees of freedom
AIC: 4131.4
```

Fig. 14: (a) Adjusted R-squared and BIC graph, (b) Summary of Model 3

iv. Evaluating the Models

The confusion matrix for the models is as follows:



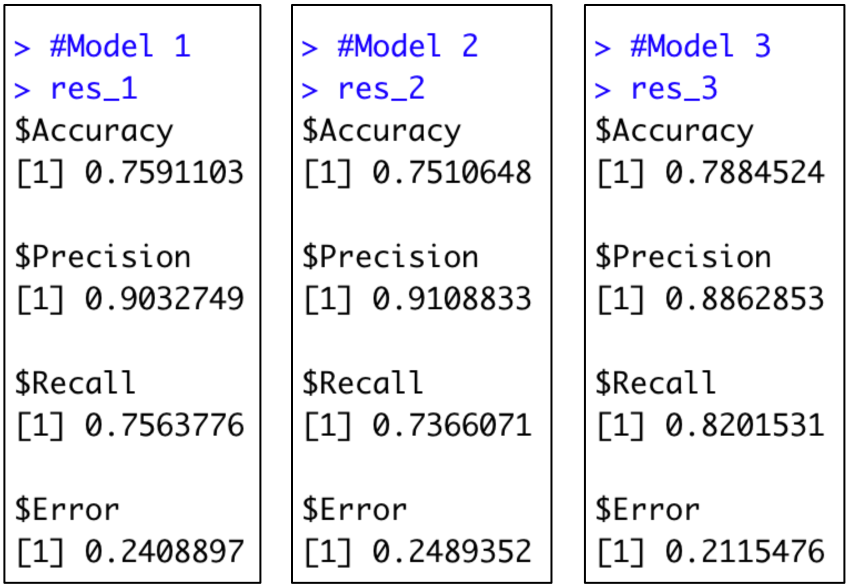
```
      predicted
actual    0    1
     0 1186  382
     1  127  418
```

```
      predicted
actual    0    1
     0 1155  413
     1  113  432
```

```
      predicted
actual    0    1
     0 1286  282
     1  165  380
```

Fig. 15: Confusion Matrix for (a) Model 1, (b) Model 2, (c) Model 3

We have calculated accuracy, precision, recall and error for all the models using confusion matrix.



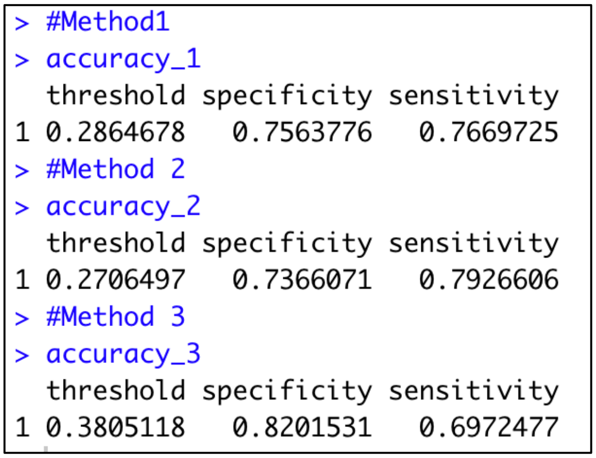
```
> #Model 1
> res_1
$Accuracy
[1] 0.7591103

$Precision
[1] 0.9032749

$Recall
[1] 0.7563776

$Error
[1] 0.2408897
```

```
> #Model 2
> res_2
$Accuracy
[1] 0.7510648

$Precision
[1] 0.9108833

$Recall
[1] 0.7366071

$Error
[1] 0.2489352
```

```
> #Model 3
> res_3
$Accuracy
[1] 0.7884524

$Precision
[1] 0.8862853

$Recall
[1] 0.8201531

$Error
[1] 0.2115476
```

Fig. 16: Accuracy, Precision, Recall and Error for (a) Model 1, (b) Model 2, (c) Model 3

The specificity and sensitivity of the models are as follows:



```
> #Method1
> accuracy_1
  threshold specificity sensitivity
1 0.2864678   0.7563776   0.7669725
> #Method 2
> accuracy_2
  threshold specificity sensitivity
1 0.2706497   0.7366071   0.7926606
> #Method 3
> accuracy_3
  threshold specificity sensitivity
1 0.3805118   0.8201531   0.6972477
```

Fig. 17: Specificity and Sensitivity for (a) Model 1, (b) Model 2, (c) Model 3

v. Selecting the Best Model

Type 1 Error (FP):
The model predicted that the customer will churn, but the customer did not actually churn.

Type 2 Error (FN):
The model predicted that the customer will not churn, but the customer actually churned.

We want our model to minimize Type 2 error. Since sensitivity can be calculated as:

$$Sensitivity = \frac{TP}{TP+FN}$$

We have selected Model 2 as our final model because it has the highest Sensitivity.

vi. Interpreting the Logistic Regression Model

1. **Internet service with fiber optic**
   Considering that all the other variables do not change, the odds that a customer churns increases by 6 times when the internet service is fiber optics rather than DSL
2. **Streaming Movies - Yes**
   Considering that all the other variables do not change, the odds that a customer churns increases by 1.84 times if the customer streams movies against when customer does not streams movies
3. **Contract – Month to month**
   Considering that all the other variables do not change, the odds that a customer churns increases by 2.05 times if the contract is month to month compared to yearly whereas the odds increases by 4.2 times when the contract is month to month compared to 2-year contract
4. **Payment Method (Electronic)**
   Considering that all the other variables do not change, the odds that a customer churns increases by approx. 1.65 times if payment is made electronically vs bank transfer/credit card/mailed cheque
5. **Tenure**
   As time the customer spent with the telco increases, the chances of the customer churning decreases

4. RESULTS

The way forward is a pyramid approach:

- Rectify the cracks and build a strong base
- Modification of the current model to suit the market dynamics
- Services/Promotions targeting reduction in churn

According to the model, the odds of the customer churning increases approximately 6 times if the customer uses Fiber Optic as Internet Service instead of DSL. To overcome this, the upper management of the Telco can come up with a promotion scheme that helps them retain these customers.

We also observed a significant increase in the odds of a customer churning if he/she is on a month-to-month contract rather than 1-year or 2-year contract. The upper management should build strategies to encourage customers to sign a 1-year or 2-year contract.

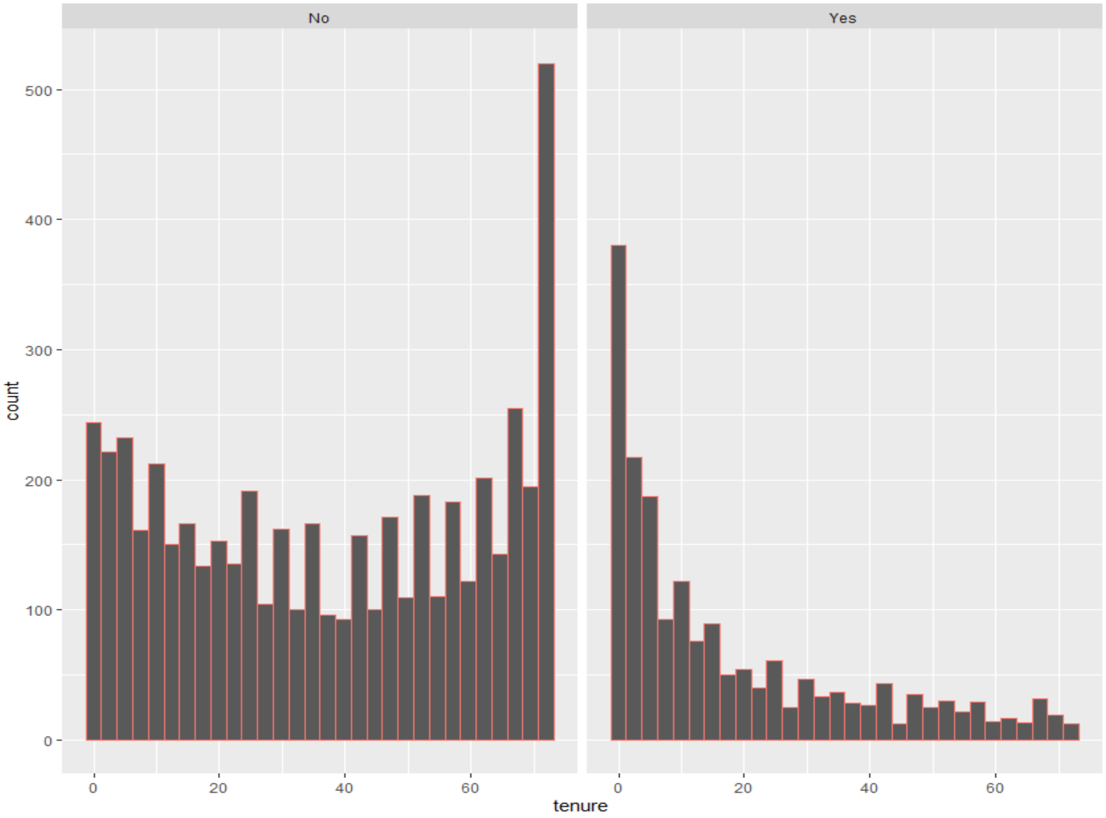


Fig. 18: (a) Histogram of retained customers and tenure, (b) Histogram of churned customers and tenure

From the graph shown in figure 18, we can see that most of the customers who churned had a tenure of less than 6 months. We would recommend the upper management to launch a loyalty bonus scheme, where customers get some concession on completing 1 year with the company.