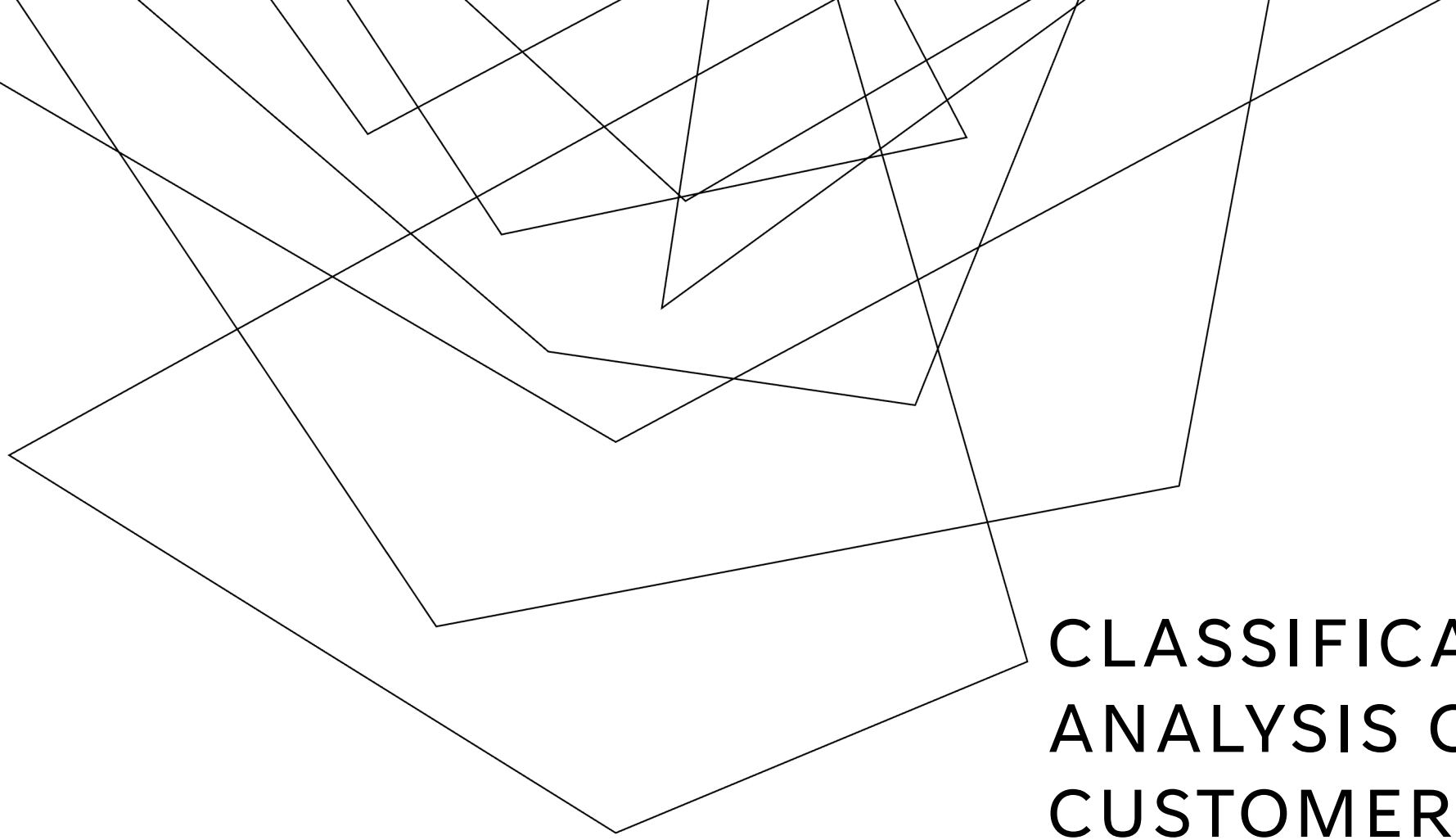
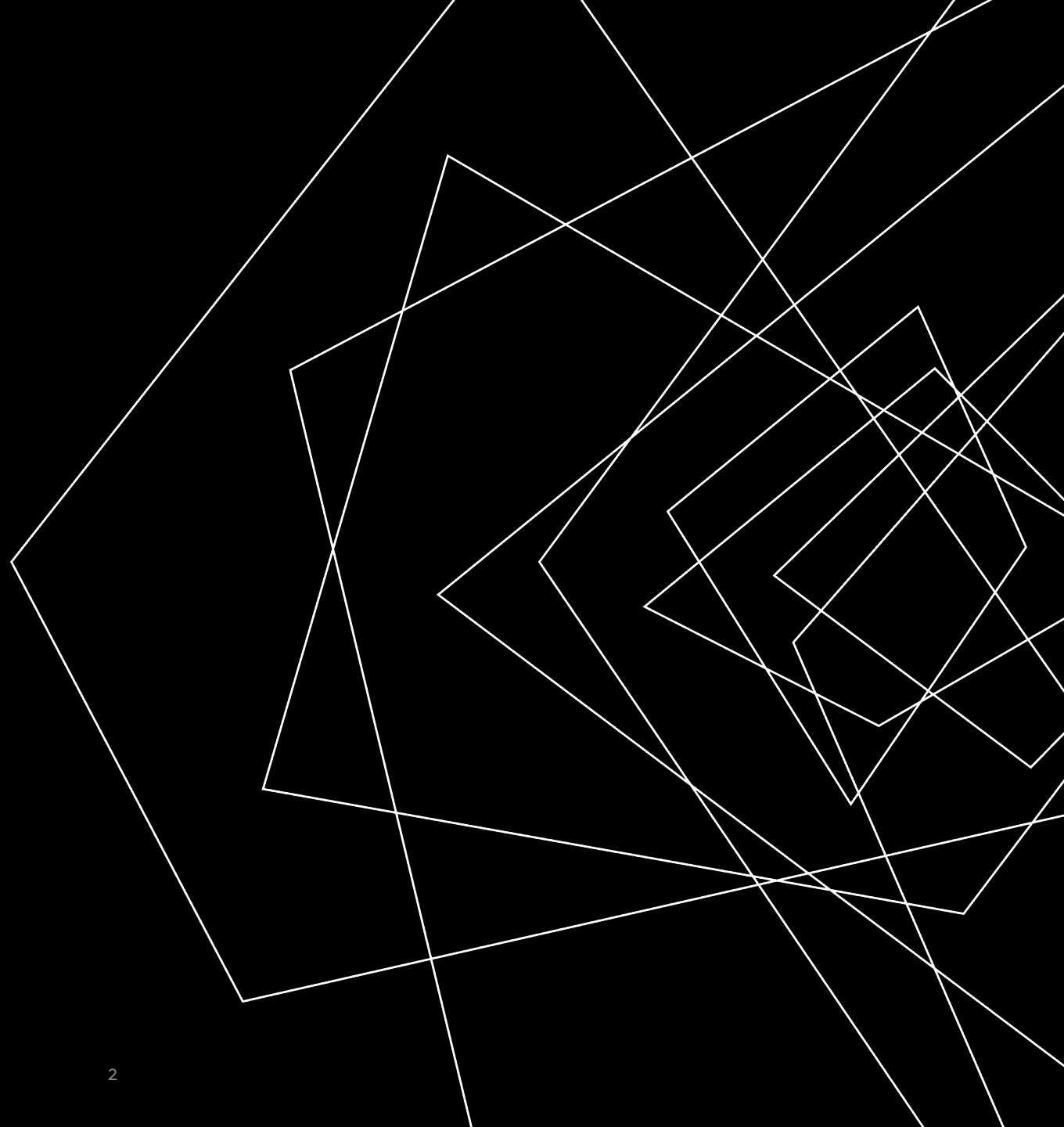


CLASSIFICATION ANALYSIS OF CUSTOMER CHURN DATA

Pranav Vinod

INTRODUCTION

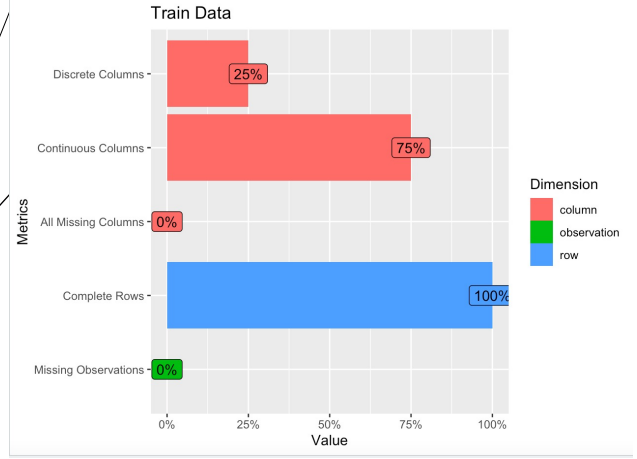
- In this project, we will try to predict whether a customer will churn, i.e. if a customer will cancel their telecom subscription based on their activity data.
- To do so, we will implement Logistic Regression, Random Forest and a Neural Network model to find the best one.
- Data Source - Link



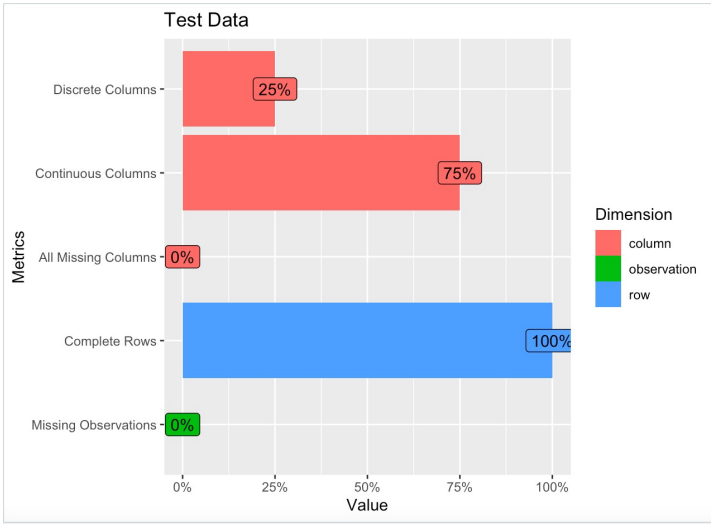
DATA DESCRIPTION

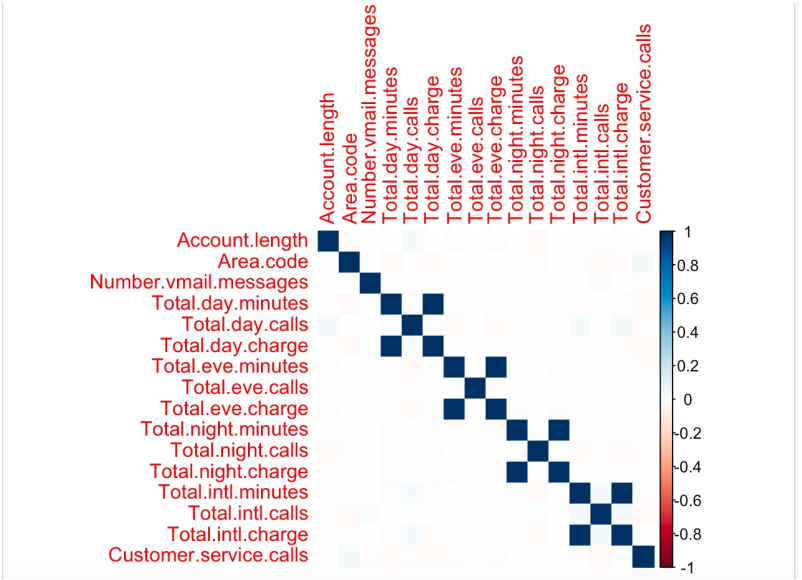
COLUMN NAME	DESCRIPTION
State	State where the account is based
Account.length	Time that the account has been active
Area.code	Area code of the account
Internation.plan	Whether account has international plan or not
Voice.mail.plan	Whether account has voice plan or not
Number.vmail.messages	Number of voice mail messages
Total.day.minutes	Total day minutes used
Total.day.calls	Day calls made
Total.day.charge	Total day charge
Total.eve.minutes	Total evening minutes used
Total.eve.calls	Evening calls made
Total.eve.charge	Total evening charge
Total.night.minutes	Total night minutes used
Total.night.calls	Night calls made
Total.night.charge	Total night charge
Total.intl.minutes	Total international minutes used
Total.intl.calls	Total international calls made
Total.intl.charge	Total international charge
Customer.service.calls	Number of service calls made
Churn	Whether the customer churns, 0 if True, 1 if False

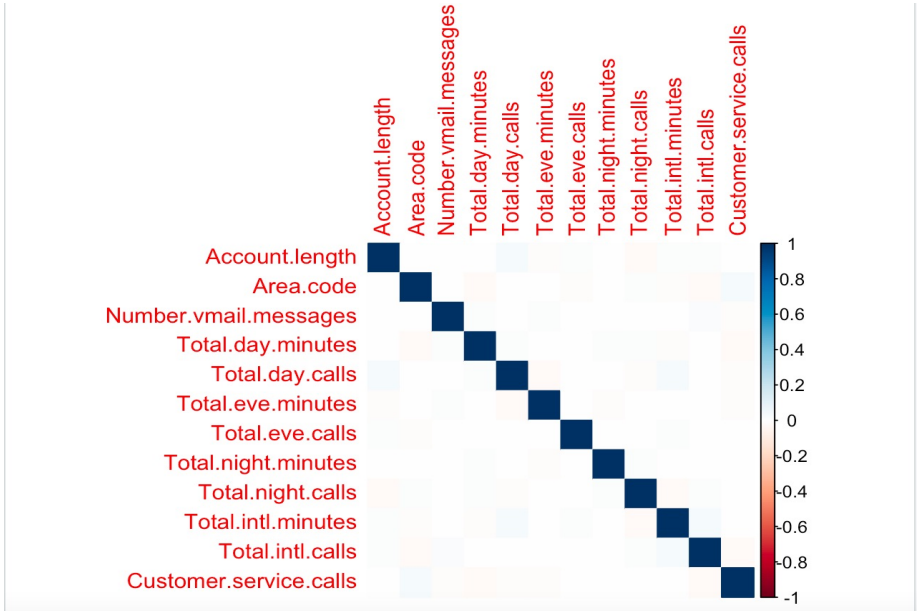
- 20 attributes : 4 categorical and 16 are numerical
- The target variable is Churn. It has 2 factor levels: True and False
- The state attribute is dropped from the dataset.



DATA EXPLORATION

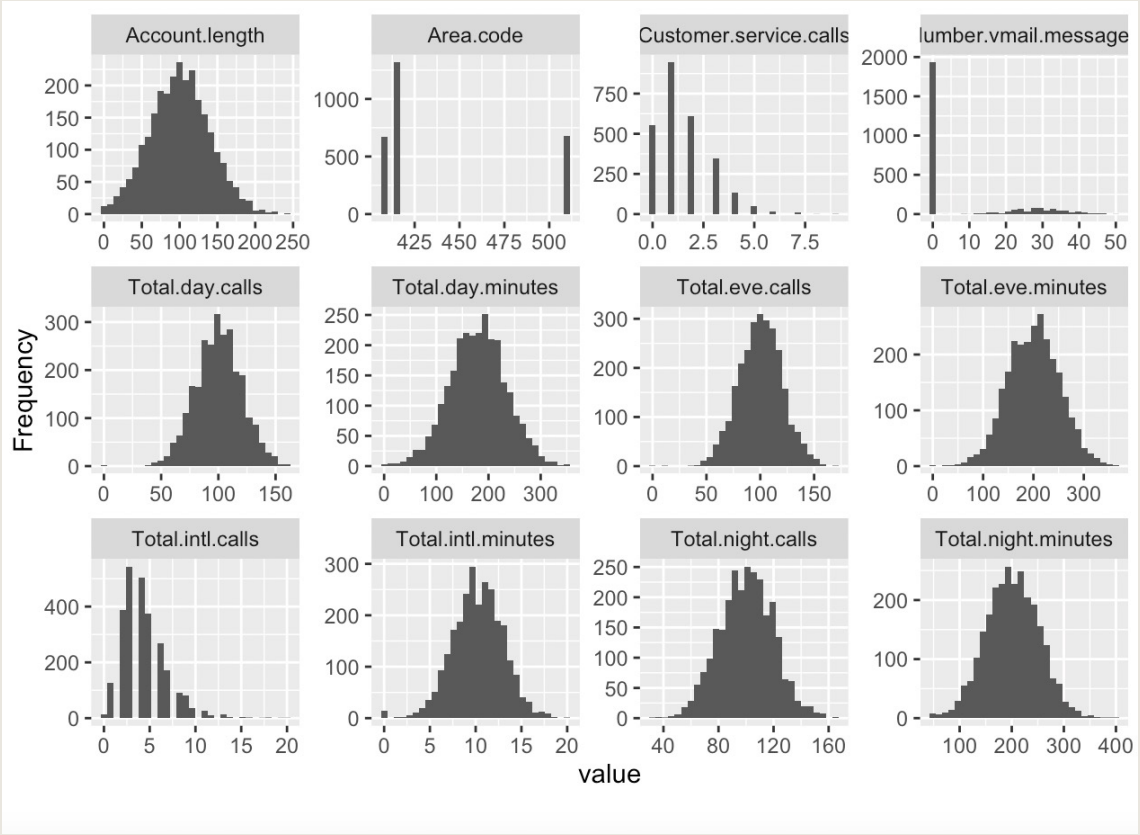




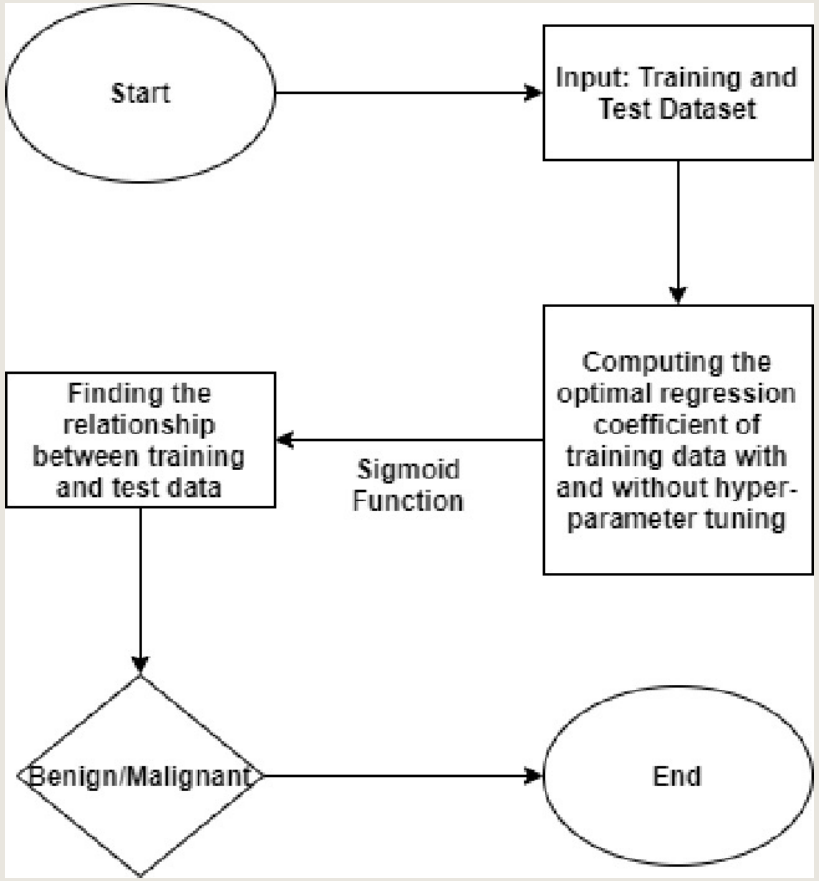


DATA EXPLORATION CONT.

```
> table(train$Churn)    # 0.172, more cases of false than true

False  True
 2278   388
> table(test$Churn)     # 0.166  more cases of false than true

False  True
  572    95
```



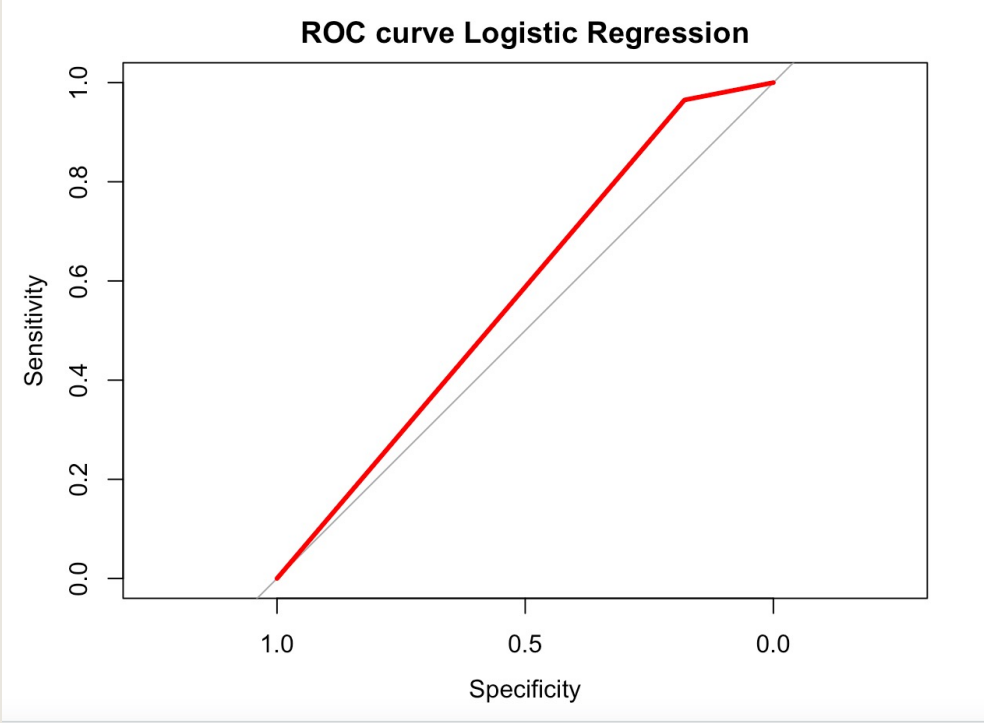
LOGISTIC REGRESSION



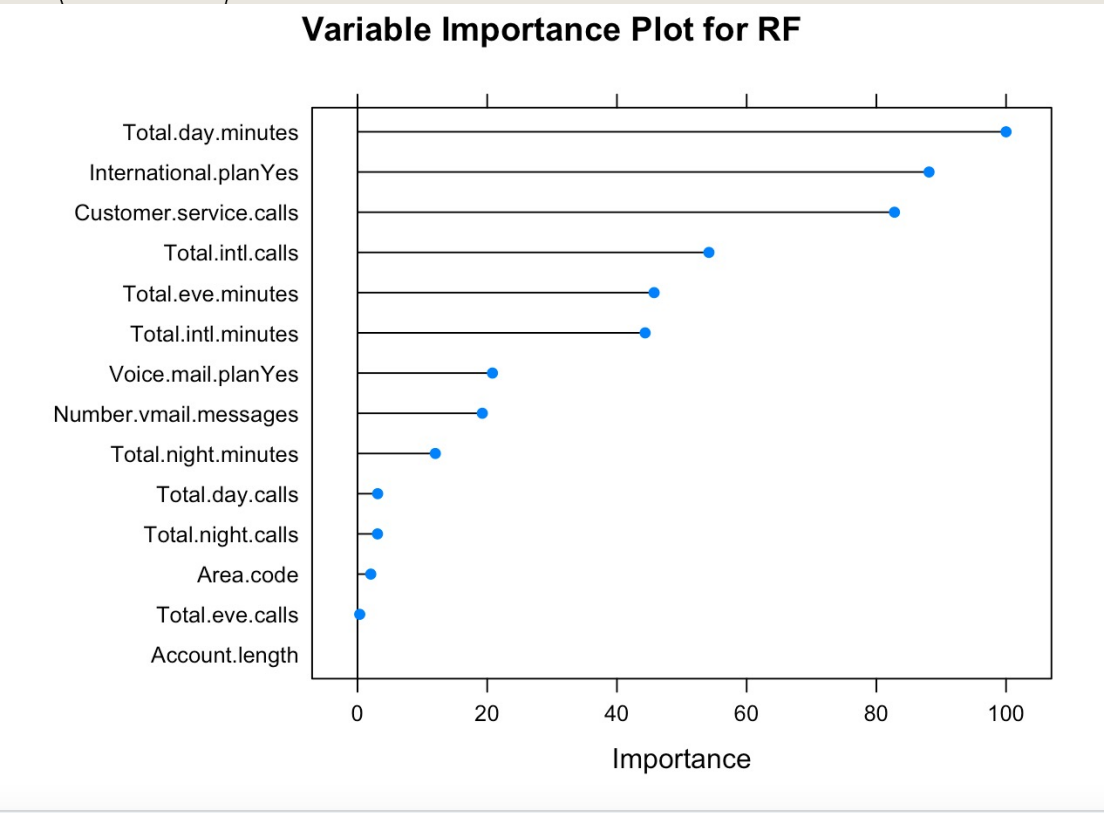

```
Call:
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = train_final)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2287   0.1982   0.3410   0.5128   2.1255

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           7.7395755  1.0361961   7.469 8.07e-14 ***
Account.length       -0.0008571  0.0015685  -0.546  0.58476
Area.code             0.0006087  0.0014690   0.414  0.67861
International.planYes -2.0977591  0.1595518 -13.148  < 2e-16 ***
Voice.mail.planYes    2.0315470  0.6553951   3.100  0.00194 **
Number.vmail.messages -0.0374140  0.0206236  -1.814  0.06966 .
Total.day.minutes     -0.0125970  0.0012187 -10.337  < 2e-16 ***
Total.day.calls       -0.0028906  0.0030978  -0.933  0.35075
Total.eve.minutes     -0.0056747  0.0012705  -4.467 7.95e-06 ***
Total.eve.calls        0.0007921  0.0030824   0.257  0.79720
Total.night.minutes   -0.0028341  0.0012418  -2.282  0.02247 *
Total.night.calls     -0.0020081  0.0031666  -0.634  0.52598
Total.intl.minutes    -0.1000652  0.0227692  -4.395 1.11e-05 ***
Total.intl.calls       0.1201464  0.0287918   4.173 3.01e-05 ***
Customer.service.calls -0.5077384  0.0440787 -11.519  < 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: 2212.2  on 2665  degrees of freedom
Residual deviance: 1730.6  on 2651  degrees of freedom
AIC: 1760.6

Number of Fisher Scoring iterations: 6
```

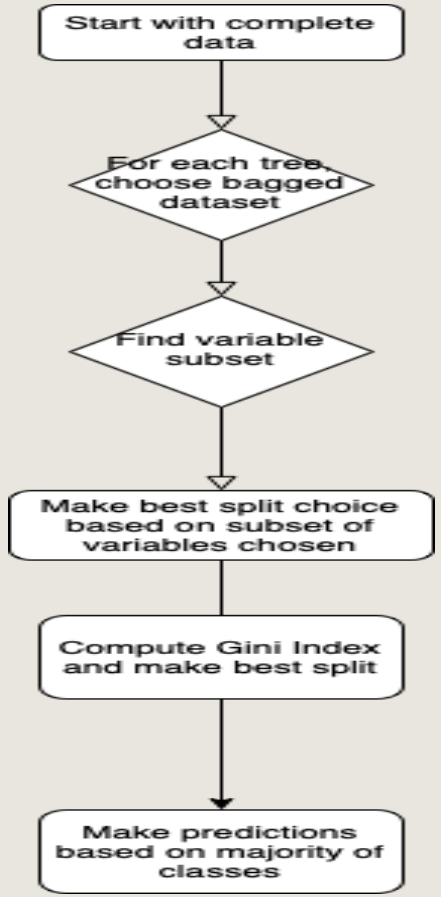
RESULTS FROM LOGISTIC REGRESSION

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  17  20
         1  78 552

               Accuracy : 0.8531
                 95% CI : (0.8239, 0.8791)
    No Information Rate : 0.8576
    P-Value [Acc > NIR] : 0.655

                  Kappa : 0.1932

 Mcnemar's Test P-Value : 8.518e-09

            Sensitivity : 0.17895
            Specificity : 0.96503
         Pos Pred Value : 0.45946
         Neg Pred Value : 0.87619
              Precision : 0.45946
                 Recall : 0.17895
                     F1 : 0.25758
             Prevalence : 0.14243
         Detection Rate : 0.02549
   Detection Prevalence : 0.05547
      Balanced Accuracy : 0.57199

       'Positive' Class : 0
```



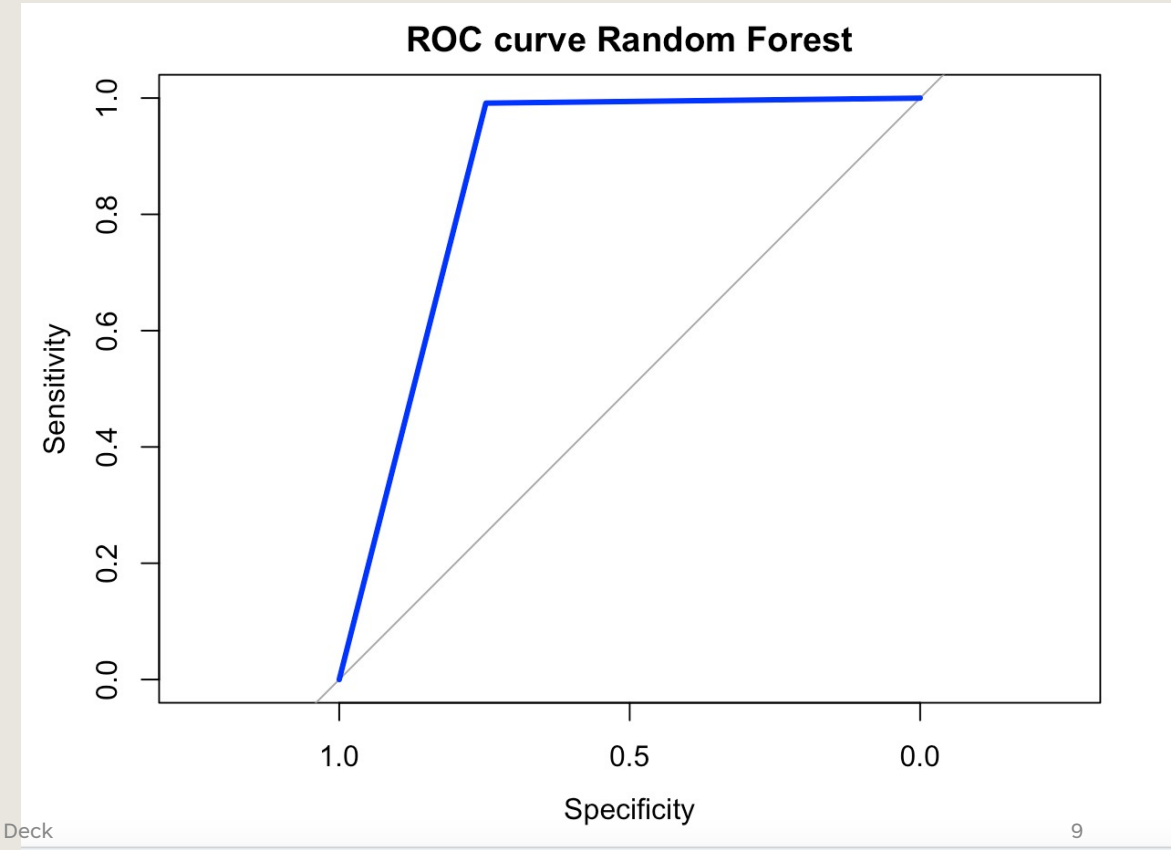
RANDOM FOREST



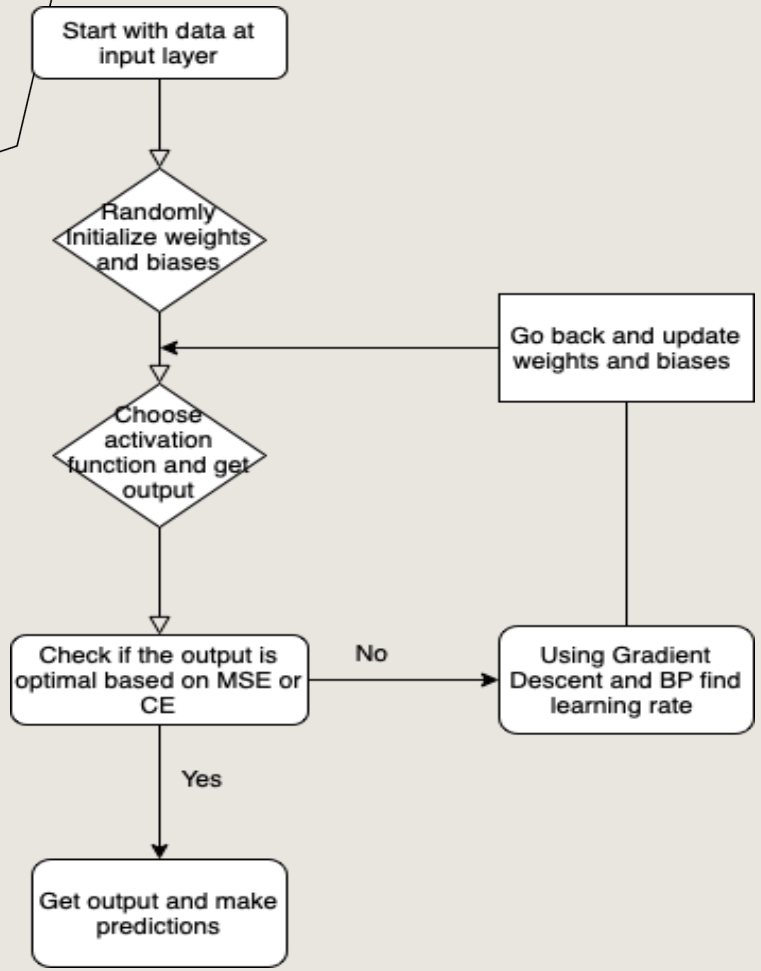


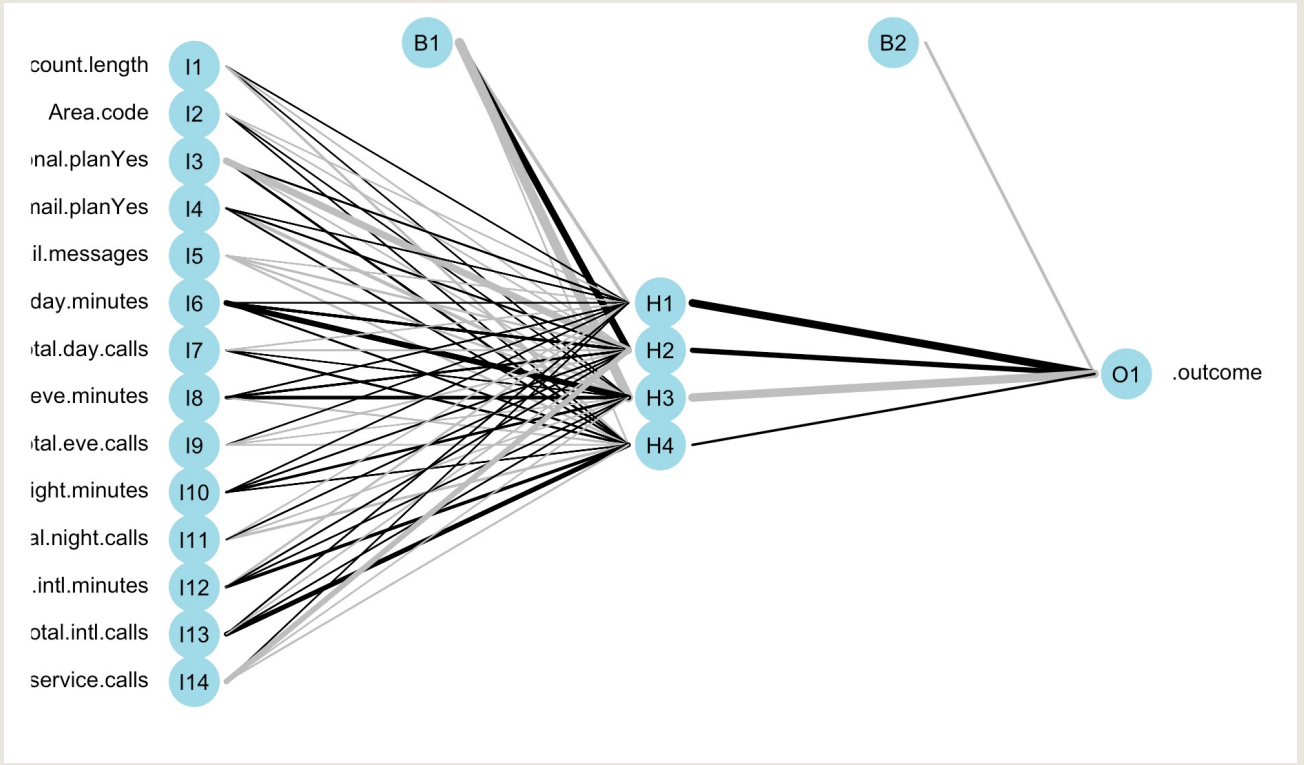
RESULTS FROM RANDOM FOREST

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  71   5
         1  24 567

               Accuracy : 0.9565
                 95% CI : (0.9382, 0.9707)
    No Information Rate : 0.8576
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8058

 Mcnemar's Test P-Value : 0.0008302

            Sensitivity : 0.7474
            Specificity : 0.9913
         Pos Pred Value : 0.9342
         Neg Pred Value : 0.9594
              Precision : 0.9342
                 Recall : 0.7474
                     F1 : 0.8304
             Prevalence : 0.1424
         Detection Rate : 0.1064
   Detection Prevalence : 0.1139
      Balanced Accuracy : 0.8693

       'Positive' Class : 0
```



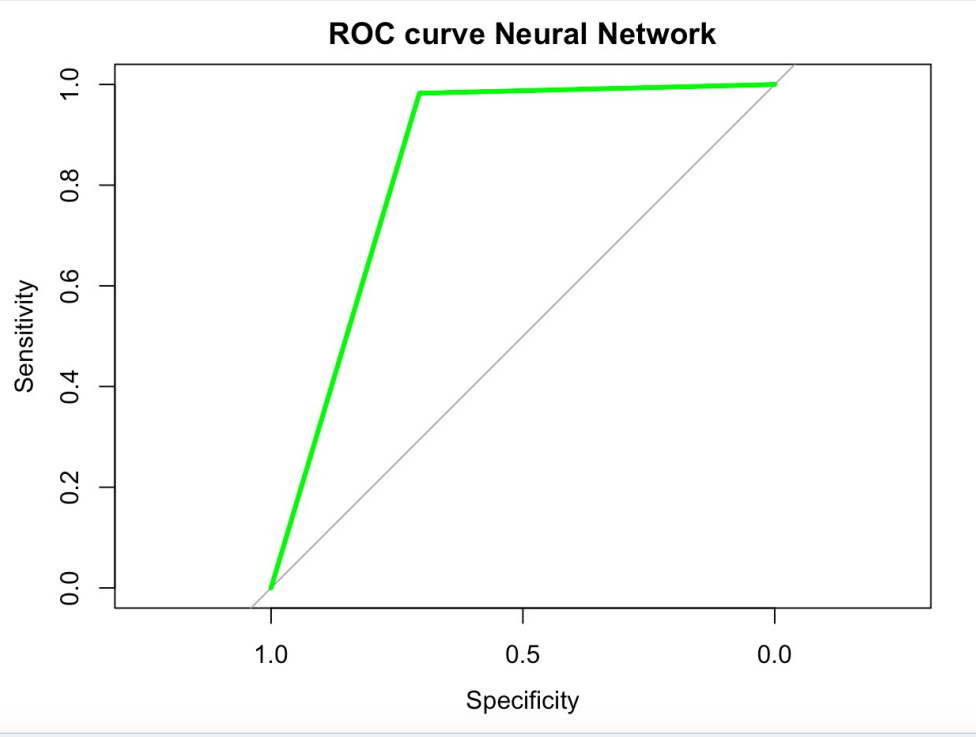
NEURAL NETWORK





RESULTS FROM NEURAL NETWORK

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  67  28
         1  10 562

               Accuracy : 0.943
                 95% CI : (0.9226, 0.9594)
    No Information Rate : 0.8846
    P-Value [Acc > NIR] : 1.793e-07

                  Kappa : 0.7468

 Mcnemar's Test P-Value : 0.00582

            Sensitivity : 0.8701
            Specificity : 0.9525
         Pos Pred Value : 0.7053
         Neg Pred Value : 0.9825
              Precision : 0.7053
                 Recall : 0.8701
                     F1 : 0.7791
             Prevalence : 0.1154
         Detection Rate : 0.1004
   Detection Prevalence : 0.1424
      Balanced Accuracy : 0.9113

       'Positive' Class : 0
```



COMPARISON B/W MODELS

	Logistic Regression	Random Forest	Neural Network
Accuracy	0.853	0.956	0.943
AUC	0.572	0.868	0.844
F1 Score	0.257	0.830	0.779

REFERENCES

1. https://www.researchgate.net/figure/Algorithmic-flow-chart-of-Logistic-Regression-Lei-et-al-2016-42-Support-vector_fig3_343472002
2. https://www.crowdanalytix.com/contests/why-customer-churn
3. Dataset
4. https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

Thank you!