**GRoup - 07**

**Tathagata Bhatta**

**Tridib Mandal**

**Liya Peter**

**Tarun Kamal Khiani**

**Biju Samuel**

**M Sarath tarun kamal khiani**

Predictive Modeling

Assignment – Customer Churn in telecom

Contents

[Goal 2](#_Toc486140338)

[Exploratory data Analysis 2](#_Toc486140339)

[Classification using logistics regression 4](#_Toc486140340)

[Overall fit 4](#_Toc486140341)

[Mcfadden r^2 5](#_Toc486140342)

[odds Ratio 5](#_Toc486140343)

[classification accuracy 5](#_Toc486140344)

[Conclusion 7](#_Toc486140345)

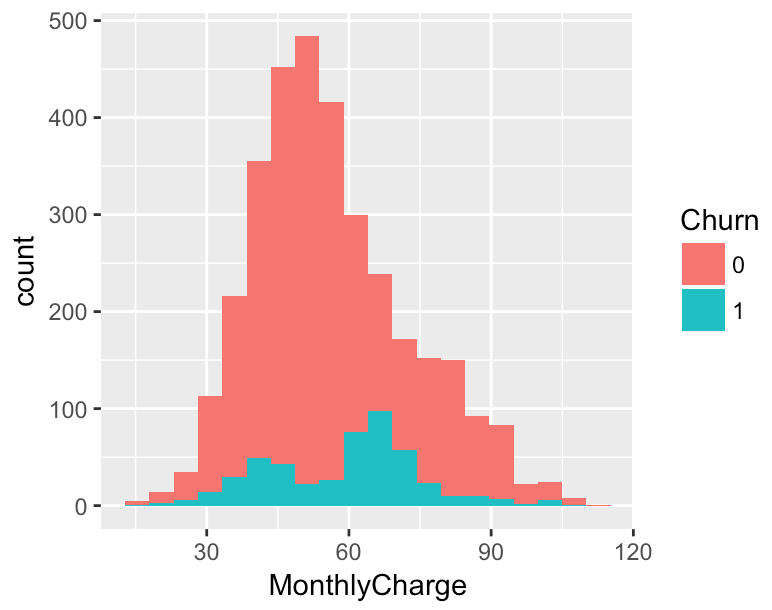
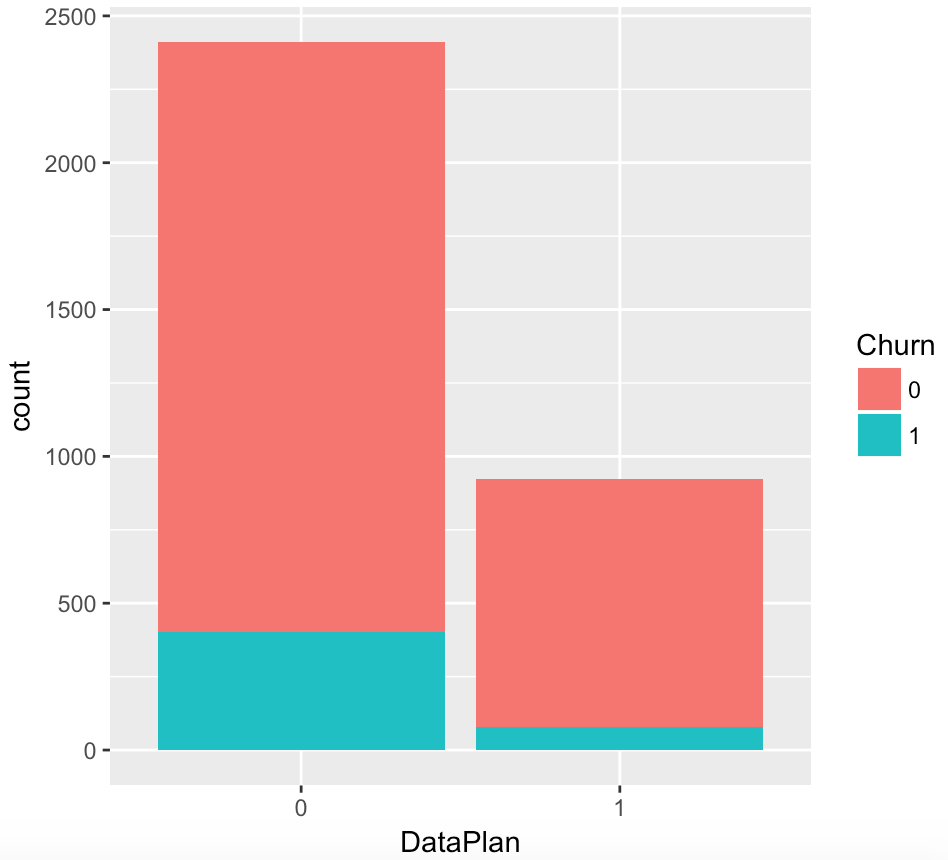
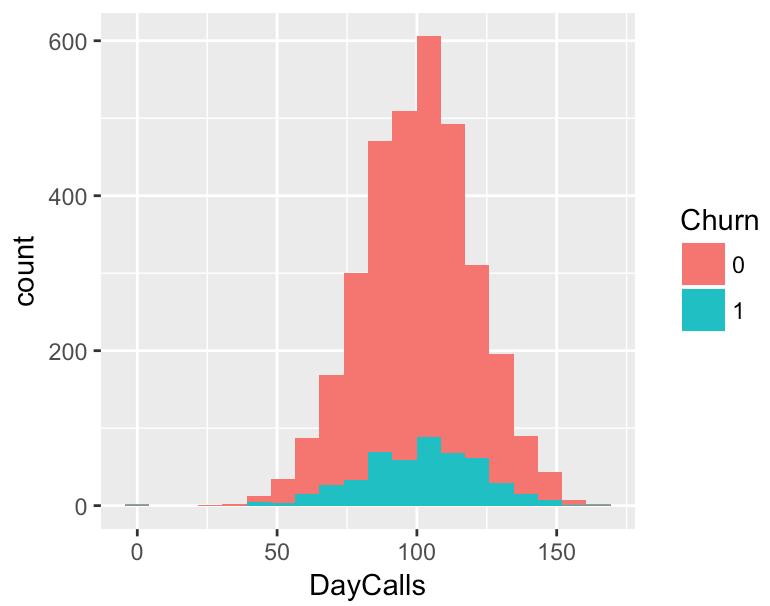
# Goal

The goal of this assignment is to build a model to predict if a customer will churn or not.

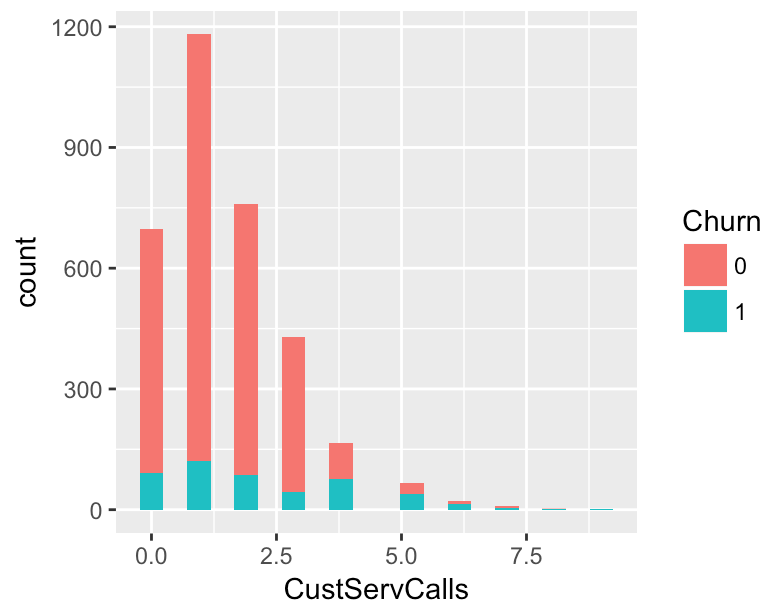
# Exploratory data Analysis

Important Observations:

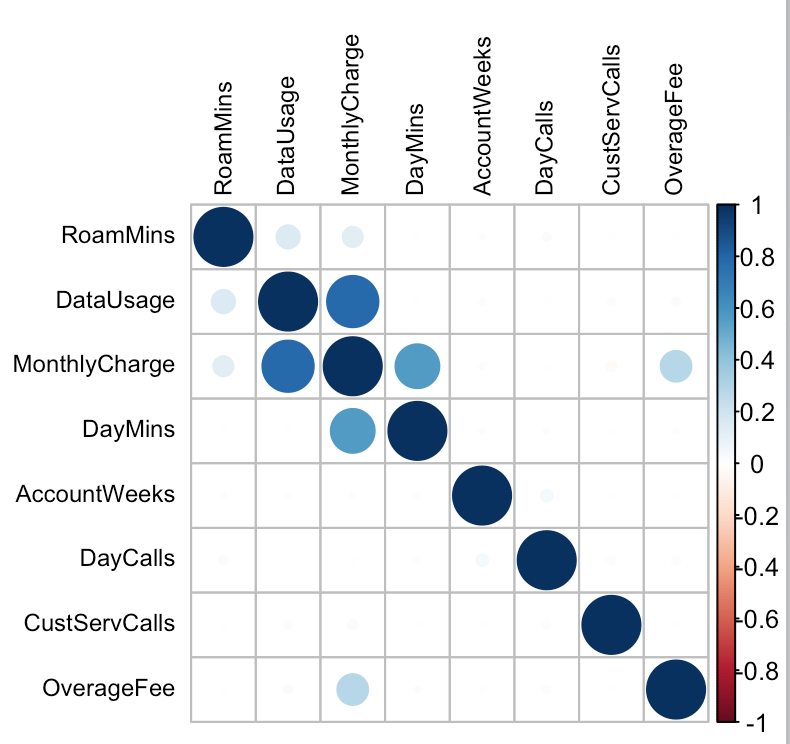
* The data quality is good as there are no null or empty fields
* Churn Rate is 14.49% hence there is no need for under and over sampling
* Churn is higher where
  + customers have no data plan
  + monthly charge is between 45 – 75
  + day calls are between 75 – 125

* Churn Rate is higher where the number of Customer Service calls are more than 3

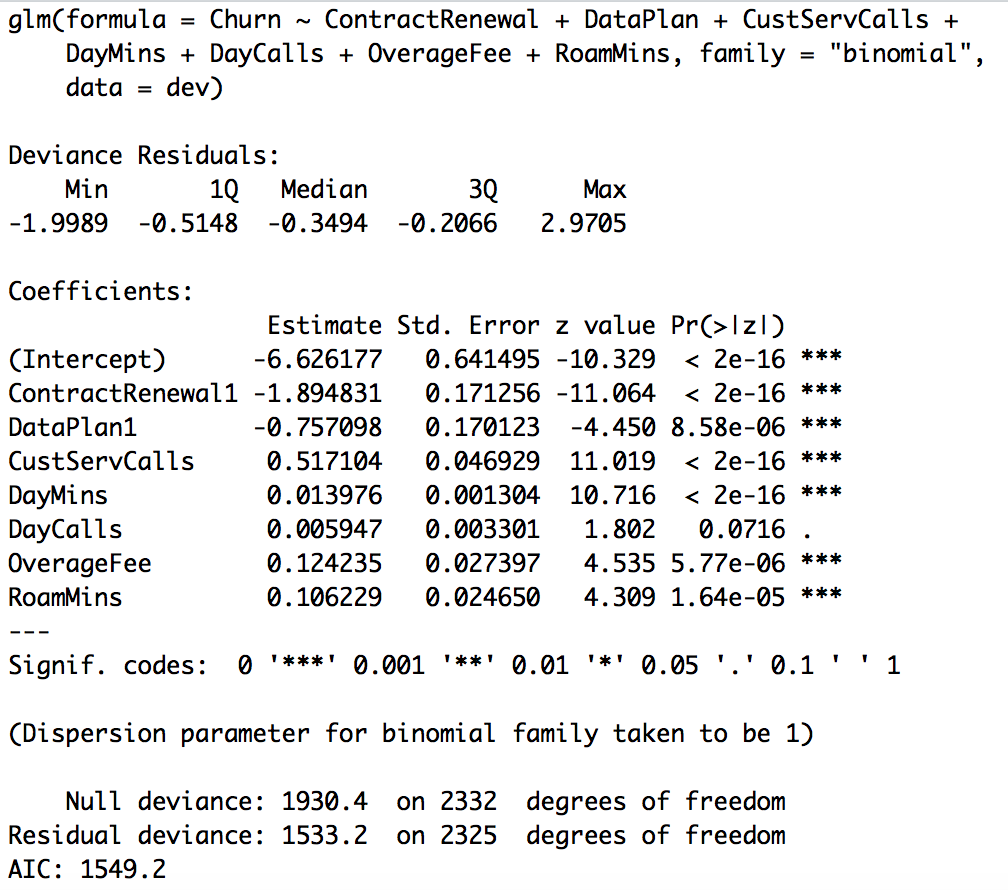


* From the correlation plot, it is evident that the following variables are correlated.
  + MonthlyCharge – DataUsage (78.0%), DayMins (57.0%)
  + Data Usage is highly correlated with Data Plan hence we can decide to use either of them in our analysis



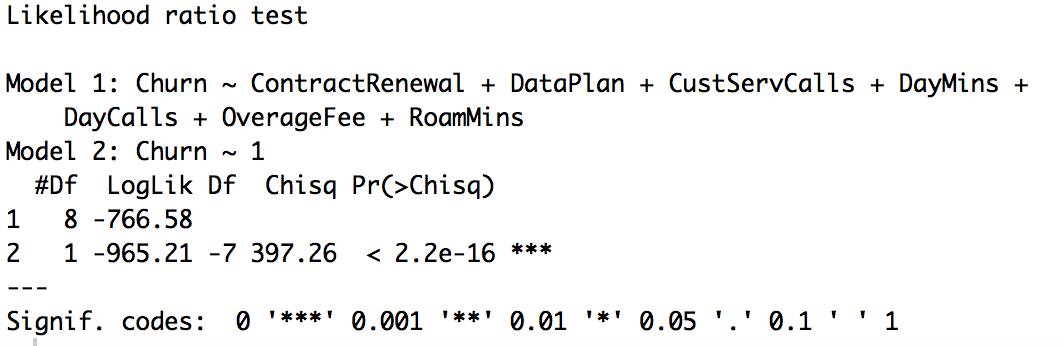
# Classification using logistics regression

Step-wise logistics regression indicates that the model performs best with the following variables – Contract Renewal, Data Plan, CustServCalls, DayMins, DayCalls, OverageFee and RoamMins. All the variables are statistically significant except **DayCalls**. Let’s explore the other parameters before arriving at a conclusion.



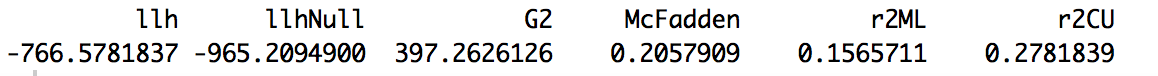
## Overall fit

The logistics regression model is a good fit (Churn depends on Contract Renewal, Data Plan, CustomerServCalls, DayMins, DayCalls, OverageFee & Roam Mins) as indicated by the likelihood ratio test.



## Mcfadden r^2

The McFadden R^2 is 0.2057909 which indicates that 20.57 % of the uncertainty of the intercept only model explains churn. A higher ratio (40 %) would be preferable but we can proceed at it indicates average performance.



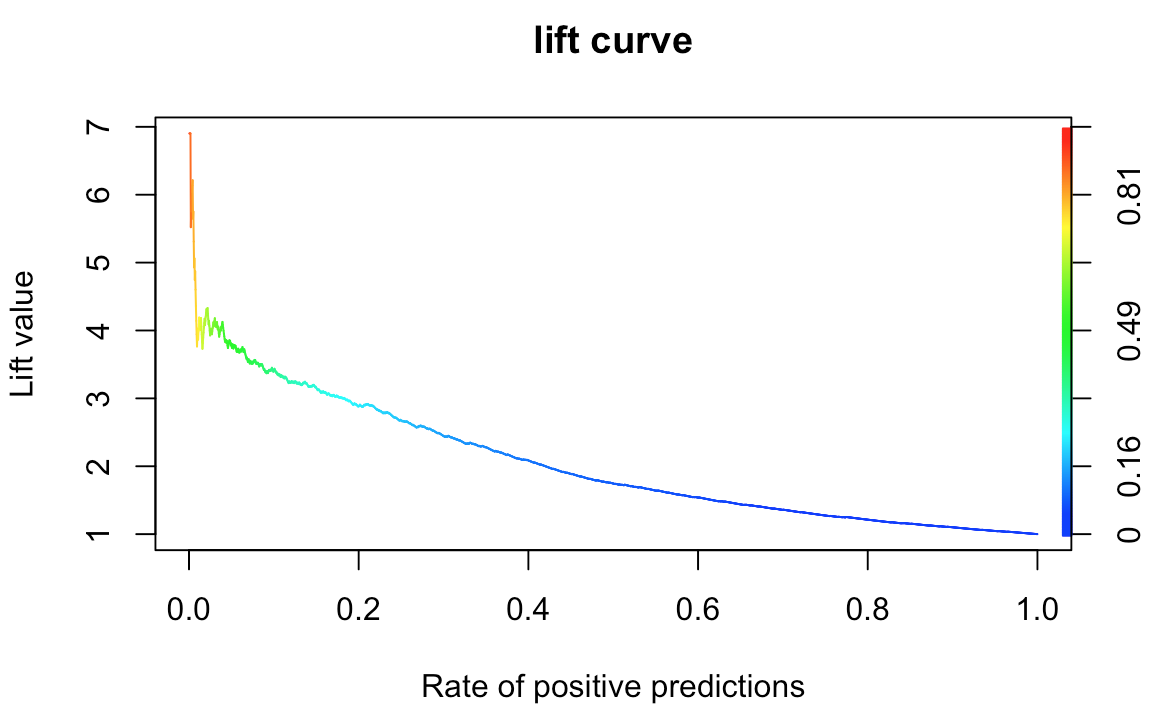
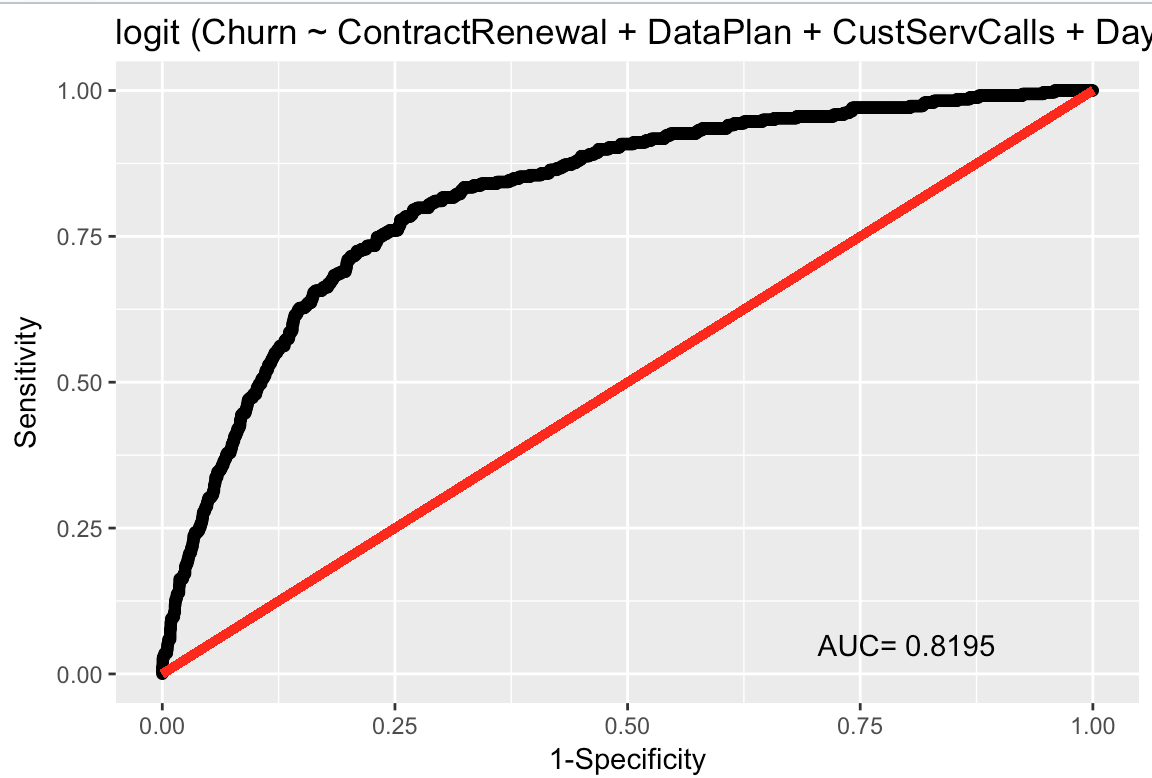
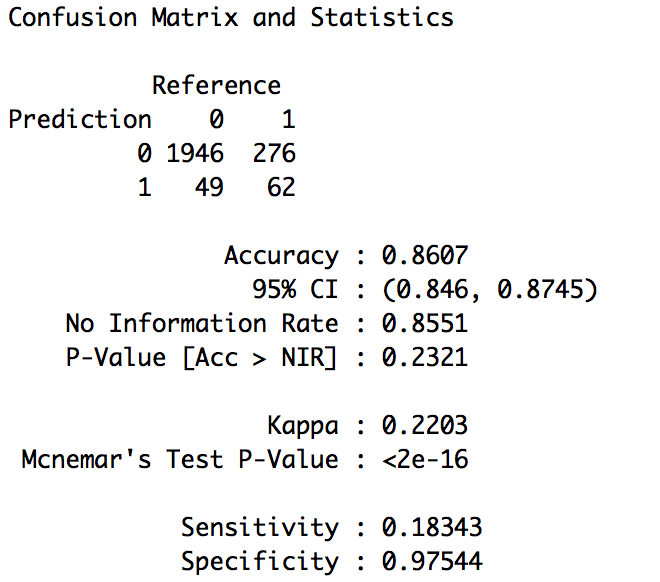
## odds Ratio

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Probability | Odds | Interpretation |
| Contract Renewal (0) | 0.8693 | 6.6514 | Customers who have not renewed their contract are 6.6514 times more likely to churn than those who have renewed their contract |
| Data Plan (0) | 0.6807 | 2.1321 | Customers who do not have a data plan are 2.1321 times more likely to churn than those who have a data plan |
| CustServCalls | 0.6264 | 1.6772 | IF the customer service calls increases by 1 call then the odds to churn increases by 1.6772 times |
| OverageFee | 0.5310 | 1.1323 | If the overage fees increases by 1 unit then the odds to churn increases by 1.1323 times |
| RoamMins | 0.5265 | 1.1121 | If the roaming mins increases by 1 unit then the odds to churn increases by 1.1121 times |
| DayMins | 0.5035 | 1.0141 | If the day mins increases by 1 unit then the odds to churn increases by 1.0141 times |
| DayCalls | 0.5014 | 1.0060 | If the customer service calls increases by 1 call then the odds to churn increases by 1.6772 times |

## classification accuracy

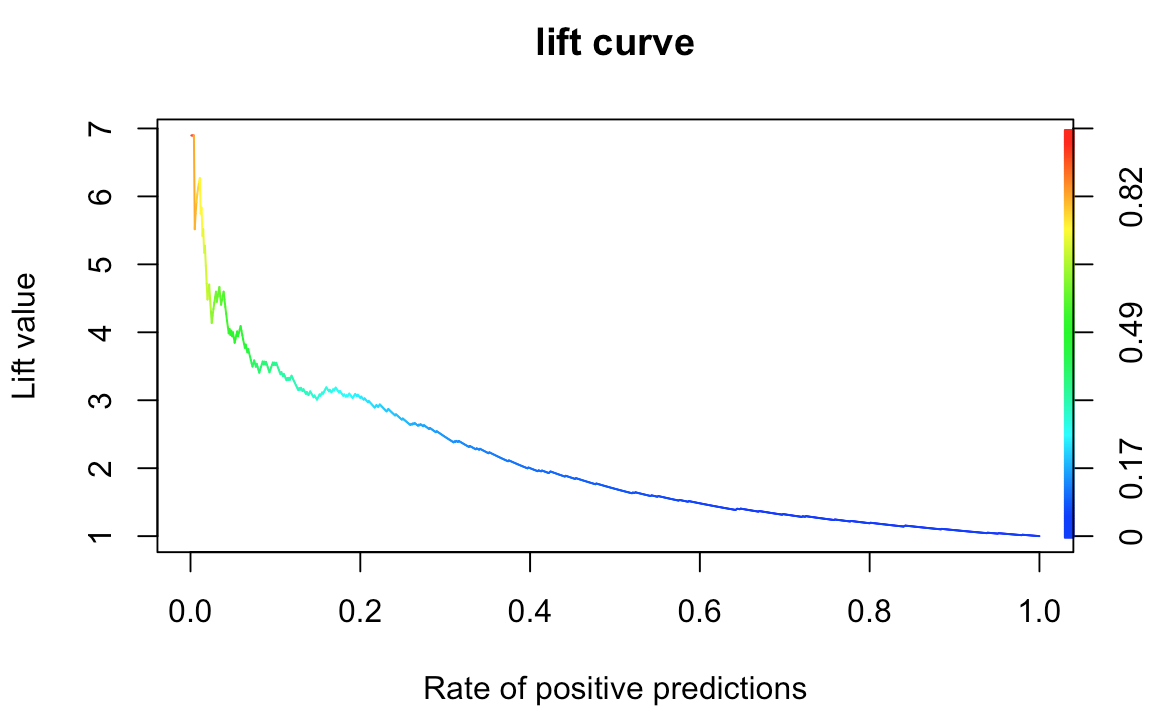
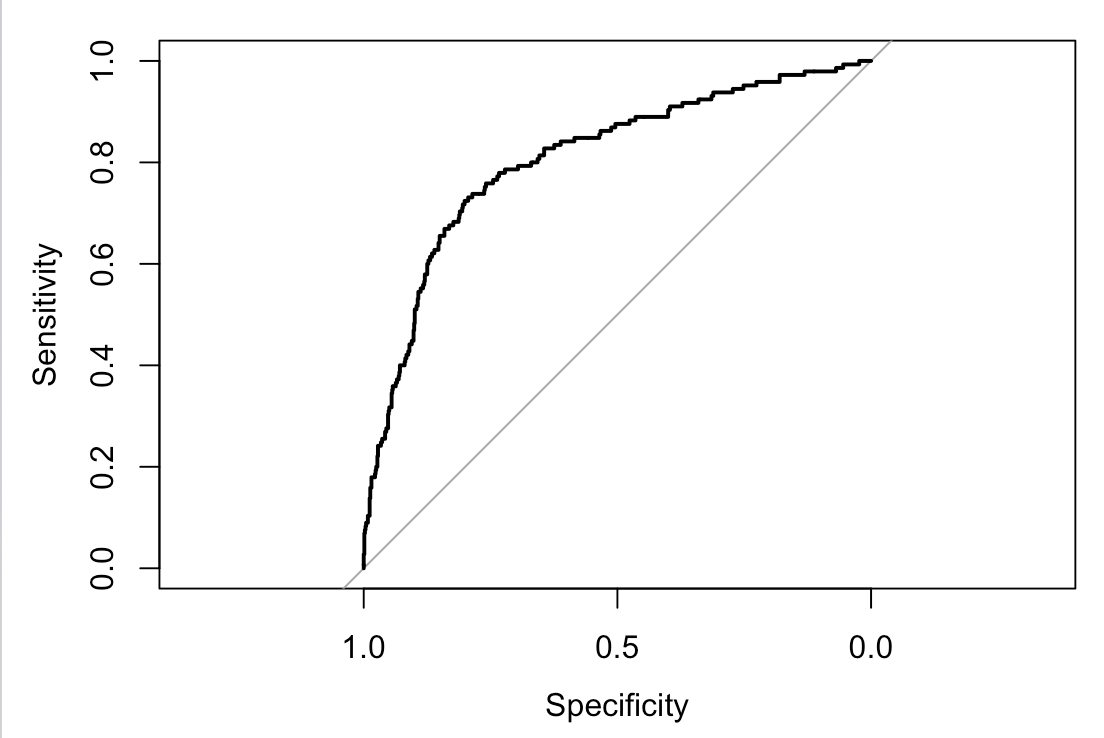
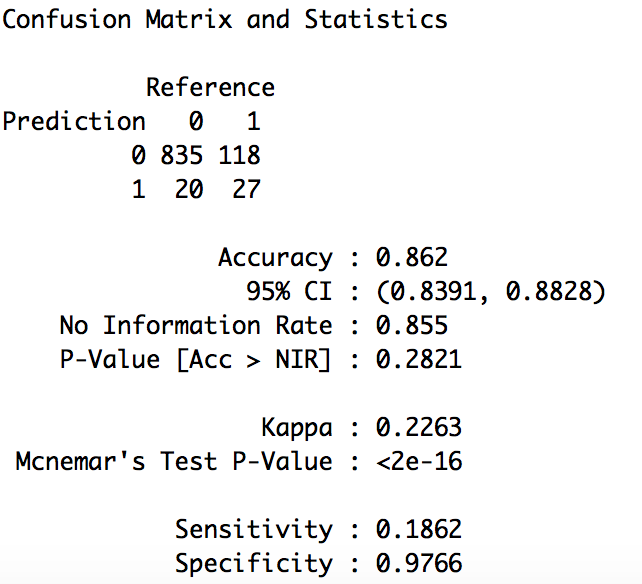
**Classification Accuracy on the development sample**

The Classification Accuracy is 86.07%, Sensitivity is 18.34%, Specificity is 97.54%, AUC is 81.95%.

****

**Classification Accuracy on the hold-out sample**

The Classification Accuracy is 86.20%, Sensitivity is 18.62%, Specificity is 97.66%, AUC is 80.65%.



# Conclusion

Need to work on a conclusion….