STAT 425
Customer Churn for Telecom Data

Himnish Jain, Emily Hylbert,
Kushal Agarwal

Agenda

- ★ Why we chose this topic
- ★ What is the dataset
- ★ How is this dataset feeling today
- ★ What we are hoping to learn from this analysis
- ★ Our findings
- ★ Conclusion

Why this topic is important

- ★ This dataset allows us the following important insights into customer behavior:
  - ○ Retention Strategies: Subscription length and churn rates are vital metrics for gauging customer loyalty.
  - ○ Strategic Pricing Decisions: Understanding the charge amount and its relationship with customer behavior can inform pricing strategy decisions.
- ★ This analysis could help the business identify the most promising targeting opportunity or next best action based on the value of a given customer.

What is the Dataset?

The dataset contains the following columns:

Column	Explanation
Call Failure	number of call failures
Complaints	binary (0: No complaint, 1: complaint)
Subscription Length	total months of subscription
Charge Amount	ordinal attribute (0: lowest amount, 9: highest amount)
Seconds of Use	total seconds of calls
Frequency of use	total number of calls
Frequency of SMS	total number of text messages
Distinct Called Numbers	total number of distinct phone calls
Age Group	ordinal attribute (1: younger age, 5: older age)
Tariff Plan	binary (1: Pay as you go, 2: contractual)
Status	binary (1: active, 2: non-active)
Age	age of customer
Customer Value	the calculated value of customer
Churn	class label (1: churn, 0: non-churn)

Key Assumptions about MLR

- ★ The residual values are normally distributed.
- ★ There must be a linear relationship between the dependent and the independent variables.
- ★ Multicollinearity: independent variables are not highly correlated with each other
- ★ The homoscedasticity assumes that the variance of the residual errors is similar across the value of each independent variable.

What are we hoping to achieve from this analysis today

★ Assist companies in understanding how to use past data to analyse and make informed decisions

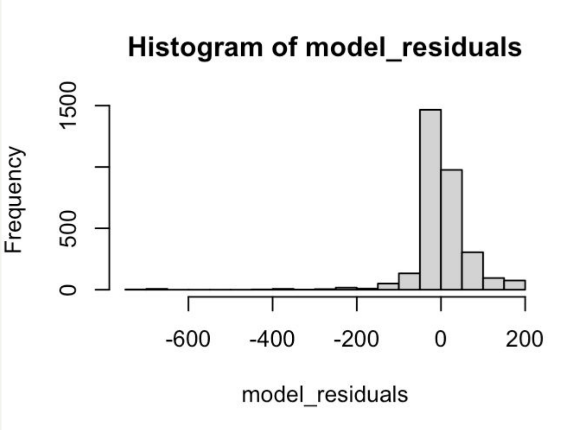
★ Understand consumer behaviour and motivations

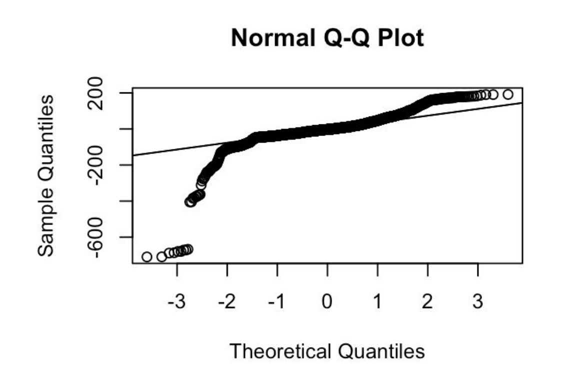
Our Findings

Distribution of Model Residuals

The histogram looks skewed to the left; hence we can not conclude the normality with enough confidence. Instead of the histogram, let's look at the residuals along the normal Q-Q plot. If there is normality, then the values should follow a straight line.

From the plot, we can observe that a few portions of the residuals lie in a straight line. Then we can assume that the residuals of the model do not follow a normal distribution.





Summary of the full model

We see that we have a strong model with High F-statistic and low p-value. But we notice some potentially collinear predictors in additional to statistically insignificant ones.

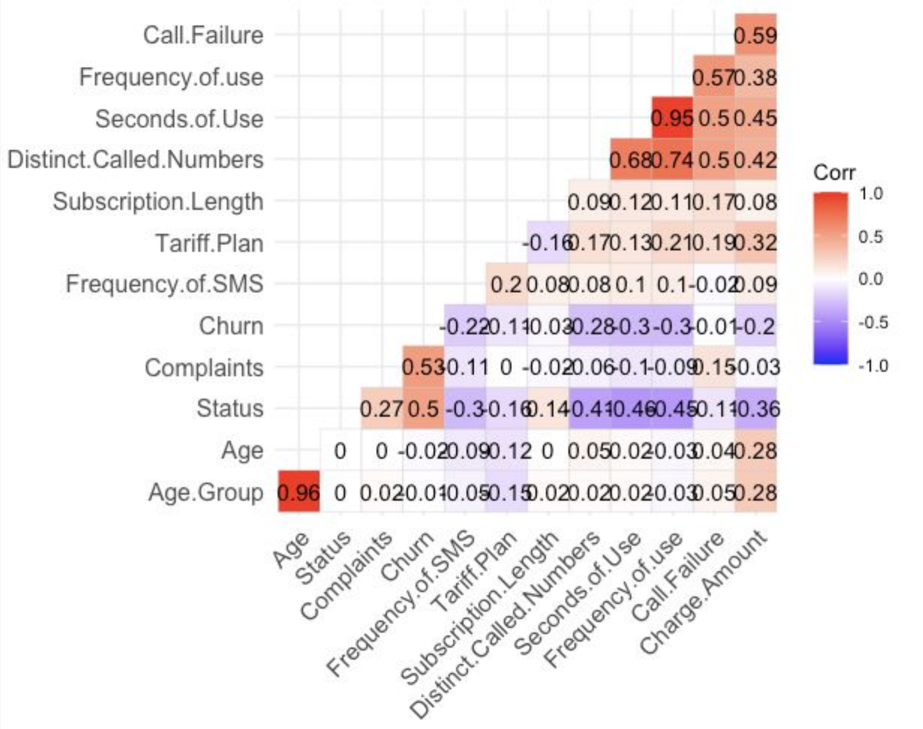
So let’s experiment with some reduced models.

```
Call:
lm(formula = Customer.Value ~ Call.Failure + Complaints + Subscription.Length +
    Charge.Amount + Seconds.of.Use + Frequency.of.use + Frequency.of.SMS +
    Distinct.Called.Numbers + Age.Group + Tariff.Plan + Status +
    Age, data = churn_data)

Residuals:
    Min      1Q  Median      3Q     Max
-709.81  -26.48   -2.63   24.24  191.43

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             160.201207  10.656561  15.033  < 2e-16 ***
Call.Failure             -0.489519   0.290861  -1.683 0.092474 .
Complaints                7.227189   5.000052   1.445 0.148439
Subscription.Length       0.741287   0.156189   4.746 2.17e-06 ***
Charge.Amount           -14.298831   1.428753 -10.008  < 2e-16 ***
Seconds.of.Use            0.047845   0.001116  42.875  < 2e-16 ***
Frequency.of.use         -0.540230   0.093055  -5.805 7.06e-09 ***
Frequency.of.SMS          4.010644   0.012108 331.234  < 2e-16 ***
Distinct.Called.Numbers   0.363675   0.112751   3.225 0.001271 **
Age.Group                -7.254712   5.211649  -1.392 0.164015
Tariff.Plan              75.507316   5.669970  13.317  < 2e-16 ***
Status                  -12.824538   3.884456  -3.302 0.000972 ***
Age                      -7.098635   0.524340 -13.538  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69.36 on 3137 degrees of freedom
Multiple R-squared:  0.9821,	Adjusted R-squared:  0.982
F-statistic: 1.432e+04 on 12 and 3137 DF,  p-value: < 2.2e-16
```

Multicollinearity

We see a strong positive correlation between age and age group and Seconds.of.use and Frequency.of.use.

Let’s build a model eliminating Age Group and Frequency.of.use due to them being lesser statistically significant than their correlated counterparts



Summary of the Reduced Model (Collinear Predictors Removed)

We see a significant improvement in F-statistic, as it goes up by 3e+03.

So a reduced model with statistically lesser significant collinear predictors are removed performs better. Let's further check this using ANOVA (F-test)

```
Call:
lm(formula = Customer.Value ~ Call.Failure + Complaints + Subscription.Length +
    Charge.Amount + Seconds.of.Use + Frequency.of.SMS + Distinct.Called.Numbers +
    Tariff.Plan + Status + Age, data = churn_data)

Residuals:
    Min      1Q  Median      3Q     Max
-705.70  -23.60   -3.71   23.56  192.33

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.653e+02  1.068e+01  15.475  < 2e-16 ***
Call.Failure           -1.377e+00  2.506e-01  -5.494 4.25e-08 ***
Complaints              9.232e+00  4.997e+00   1.848   0.0648 .
Subscription.Length     7.313e-01  1.570e-01   4.659 3.31e-06 ***
Charge.Amount          -1.008e+01  1.217e+00  -8.288  < 2e-16 ***
Seconds.of.Use          4.186e-02  4.427e-04  94.540  < 2e-16 ***
Frequency.of.SMS        4.010e+00  1.201e-02 333.854  < 2e-16 ***
Distinct.Called.Numbers 1.461e-01  1.041e-01   1.403   0.1606
Tariff.Plan             6.532e+01  5.257e+00  12.424  < 2e-16 ***
Status                 -8.420e+00  3.830e+00  -2.199   0.0280 *
Age                    -7.832e+00  1.548e-01 -50.578  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69.74 on 3139 degrees of freedom
Multiple R-squared:  0.9819,    Adjusted R-squared:  0.9818
F-statistic: 1.699e+04 on 10 and 3139 DF,  p-value: < 2.2e-16
```

ANOVA of the Full model and the first reduced model

ANOVA tells us a different story. The p-value is much lesser than 0.05, rejecting the null hypothesis that the reduced model is better.

We find this interesting because it shows that the full model with collinear variables might explain lesser variance, but it is a better fit.

```
> anova(cust_value_model, colinearity_model)
Analysis of Variance Table

Model 1: Customer.Value ~ Call.Failure + Complaints + Subscription.Length +
    Charge.Amount + Seconds.of.Use + Frequency.of.use + Frequency.of.SMS +
    Distinct.Called.Numbers + Age.Group + Tariff.Plan + Status +
    Age
Model 2: Customer.Value ~ Call.Failure + Complaints + Subscription.Length +
    Charge.Amount + Seconds.of.Use + Frequency.of.SMS + Distinct.Called.Numbers +
    Tariff.Plan + Status + Age
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1   3137 15092825
2   3139 15268753 -2   -175929 18.283 1.275e-08 ***
```

Summary of The Model with Only statistically Significant predictors

The F-statistic tests the overall significance of the model by comparing the variability explained by the model to the variability not explained. The extremely low p-value (< 2.2e-16) indicates that the model as a whole is highly significant in predicting Customer.Value.

```
Call:
lm(formula = Customer.Value ~ Subscription.Length + Charge.Amount +
    Seconds.of.Use + Frequency.of.use + Frequency.of.SMS + Distinct.Called.Numbers
+
    Tariff.Plan + Status + Age, data = churn_data)

Residuals:
    Min      1Q  Median      3Q     Max
-706.92  -25.94   -4.10   23.20  194.26

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             159.600223  10.646945  14.990  < 2e-16 ***
Subscription.Length       0.706071   0.155474   4.541 5.80e-06 ***
Charge.Amount           -15.935690   1.103039 -14.447  < 2e-16 ***
Seconds.of.Use            0.048607   0.001022  47.543  < 2e-16 ***
Frequency.of.use         -0.625734   0.079631  -7.858 5.32e-15 ***
Frequency.of.SMS          4.008469   0.011945 335.584  < 2e-16 ***
Distinct.Called.Numbers   0.390373   0.111383   3.505 0.000463 ***
Tariff.Plan              78.641296   5.490352  14.324  < 2e-16 ***
Status                  -13.845853   3.594606  -3.852 0.000120 ***
Age                      -7.757874   0.152254 -50.954  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 69.39 on 3140 degrees of freedom
Multiple R-squared:  0.982,	Adjusted R-squared:  0.982
F-statistic: 1.907e+04 on 9 and 3140 DF,  p-value: < 2.2e-16
```

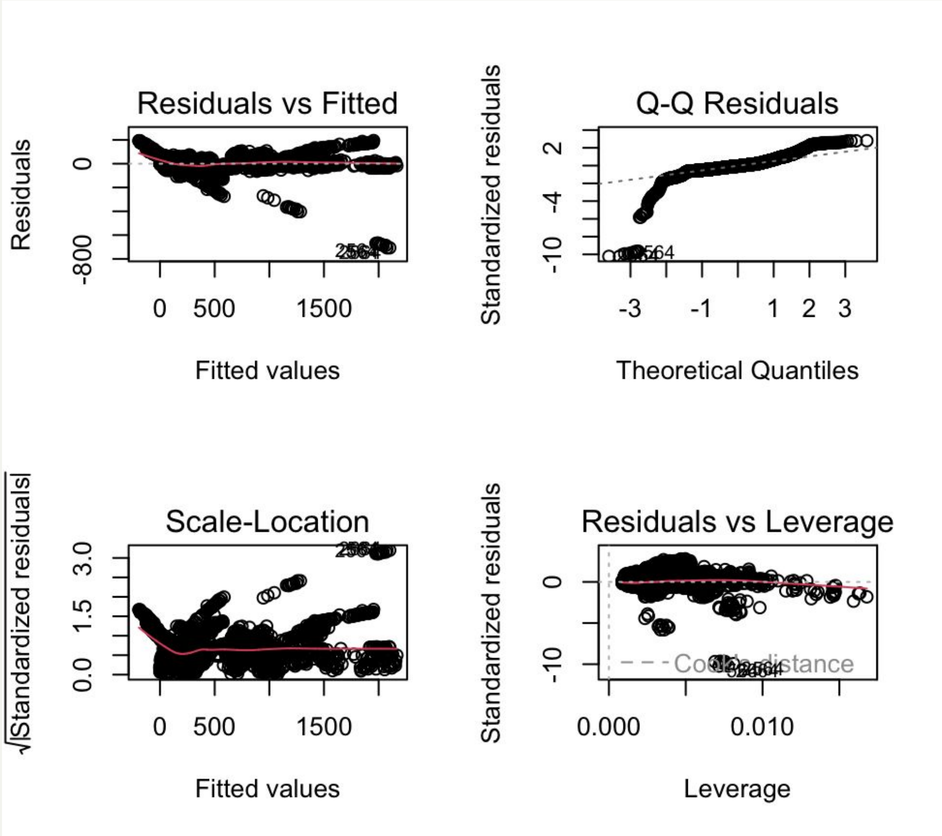
ANOVA of the Full model and The 2nd reduced model

The p-value associated with the F test is 0.1201 which is greater than 0.05. Therefore, we fail to reject the null hypothesis that the reduced model is significantly different from the full model. This suggests that reduced model provides a better fit.

```
> anova(cust_value_model, significant_only_model)
Analysis of Variance Table

Model 1: Customer.Value ~ Call.Failure + Complaints + Subscription.Length +
    Charge.Amount + Seconds.of.Use + Frequency.of.use + Frequency.of.SMS +
    Distinct.Called.Numbers + Age.Group + Tariff.Plan + Status +
    Age
Model 2: Customer.Value ~ Subscription.Length + Charge.Amount + Seconds.of.Use +
    Frequency.of.use + Frequency.of.SMS + Distinct.Called.Numbers +
    Tariff.Plan + Status + Age
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1   3137 15092825
2   3140 15120908 -3    -28083 1.9457 0.1201
```

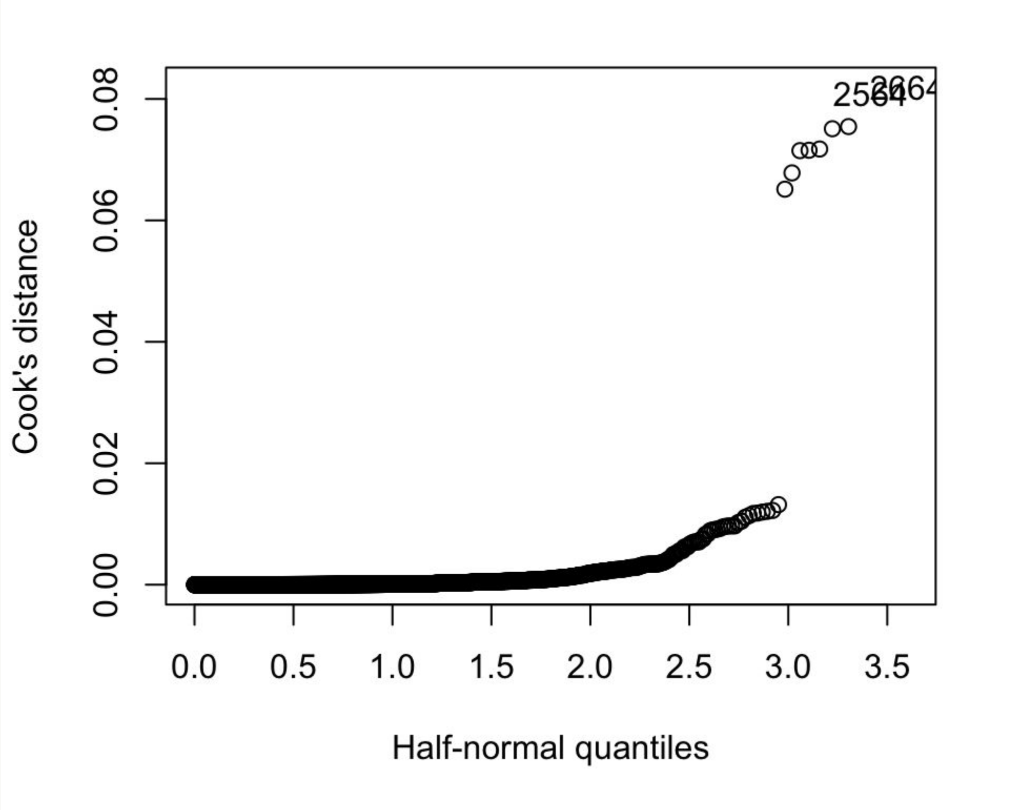
Plots - 2nd Reduced Model

- ★ Residuals vs. Fitted
  - ○ Displays Linearity
- ★ Q-Q Residuals
  - ○ Demonstrates data is not normally distributed
- ★ Scale-Location
  - ○ Variance of residuals mostly constant
- ★ Residuals vs. Leverage
  - ○ Points further off from 0, potential for influential points



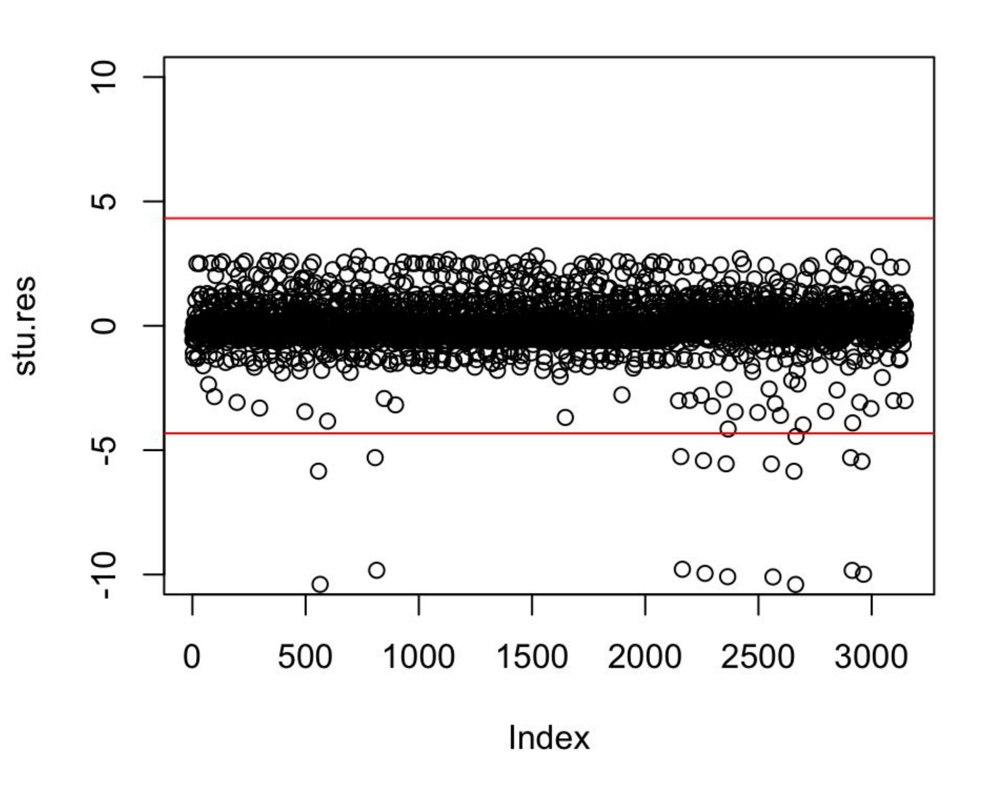
Influential Points

- ★ The Cook's Distance for each observation is less than 1, meaning there are **no influential values,** but this further exhibits the presence of outliers



Outliers

★ We found many outliers within the data. Removing them and basing a model off of the clean dataset would improve the output.



Predicting Churn

Interpretation

We got a prediction accuracy of 89.4% which is helpful in identifying core focus areas.

The model shows good predictive ability with high accuracy and sensitivity. However, specificity and the balance between false positives and negatives suggest room for improvement, especially in correctly predicting churn cases. The significant predictors, such as `Complaints`, `Call.Failure`, and `Charge.Amount`, highlight areas potentially impacting customer decisions to churn. This data can help us and many companies identify and form strategies around customer retention.

```
          Reference
Prediction   0   1
         0 511  54
         1  13  52

               Accuracy : 0.8937
                 95% CI : (0.8669, 0.9166)
    No Information Rate : 0.8317
    P-Value [Acc > NIR] : 7.565e-06
```

Conclusions

- ★ Based on our analysis, this company has the ability to improve customer value through focusing on more significant predictors, including charge amount, age, complaints, and minutes of use.
  - Focusing advertisements towards specific payment plans and age groups as well as addressing issues found in complaints can yield a higher customer value and lower churn rate.
- ★ Using our prediction model, the company can very accurately predict the amount of customers they expect to maintain, but the model is less accurate for customers who will not return.

Questions?