Customer churn analysis: introduction

(i) Churn analysis is the evaluation of a company's customer loss rate in order to reduce it.

Customer churn analysis: Telco dataset

"Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs." [IBM Sample Data Sets] Content. Each row represents a customer, each column contains customer's attributes described on the column Metadata. The data set includes information about:

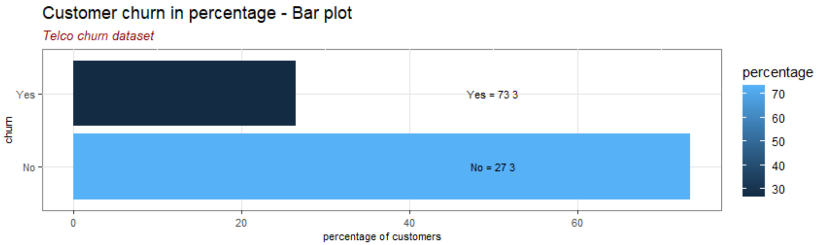
(i) Customers who left within the last month – the column is called Churn

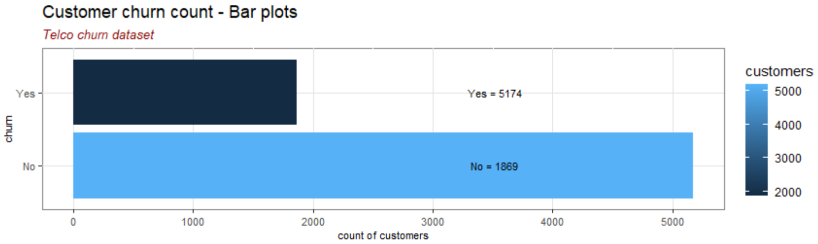
(ii) Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies

(iii) Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges

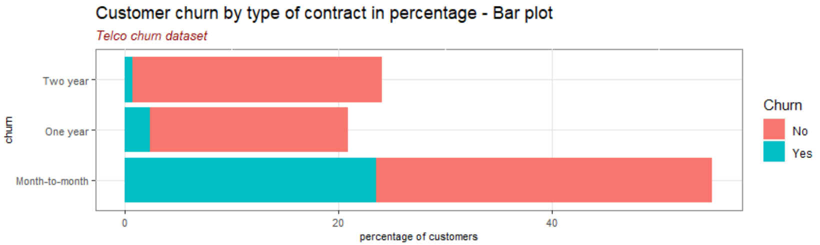
(iv) Demographic info about customers – gender, age range, and if they have partners and dependents

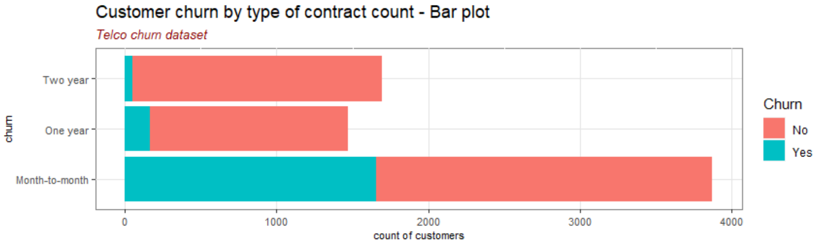
Visualization: churn rate



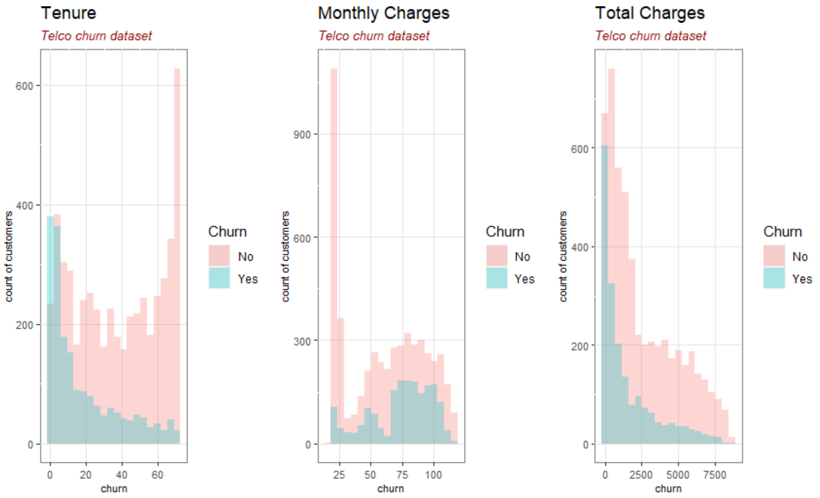


Visualization: churn rate by type of contract

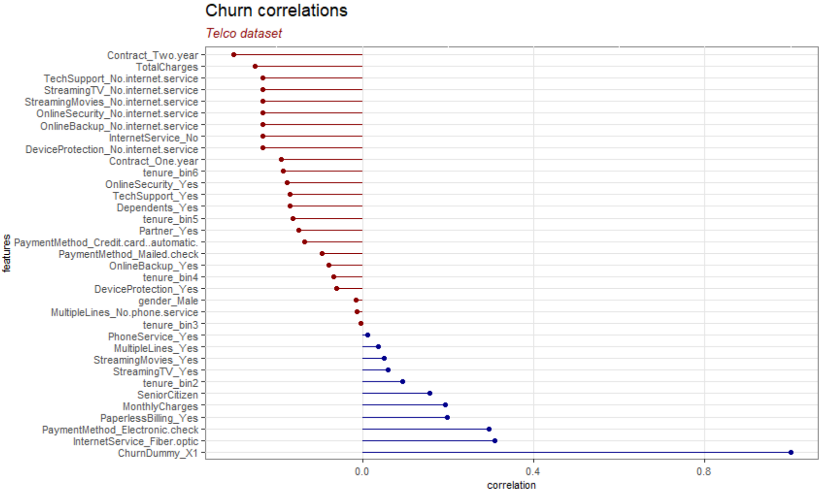




Visualization: churn rate by type of fidelity and charges



Visualization: predictors correlated to churn



Modeling: Logistic regression (1/5)

Consider n independent observations $y_1, ..., y_n$ for which we assume a Bernoulli distribution conditionally on a set of p categorical or numerical covariates x_j, for $j = 1, ..., p$. The model is given by

$$g\Big(E[y_i \mid \mathbf{x}_i]\Big) = g\big(\pi_i\big) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = \mathbf{x}_i^T \boldsymbol{\beta}$$

with $i = 1, ..., n$, with $\mathbf{x}_i^T = (1, x_{i1}, ..., x_{ip})^T$ and $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$.

The canonical link function is the logistic link. The purpose of a link function is to link the linear predictors to the mean of the response. The logistic link ensures that the predicted values lie in the interval $[0, 1]$. We have

$$g\Big(E\big[y_i \mid \mathbf{x}_i\big]\Big) = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}$$

Other link functions $g(.)$ exist, such as the probit link or the complementary log-log link. It follows naturally that

$$E[y_i \mid \mathbf{x}_i] = \pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}} = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$$

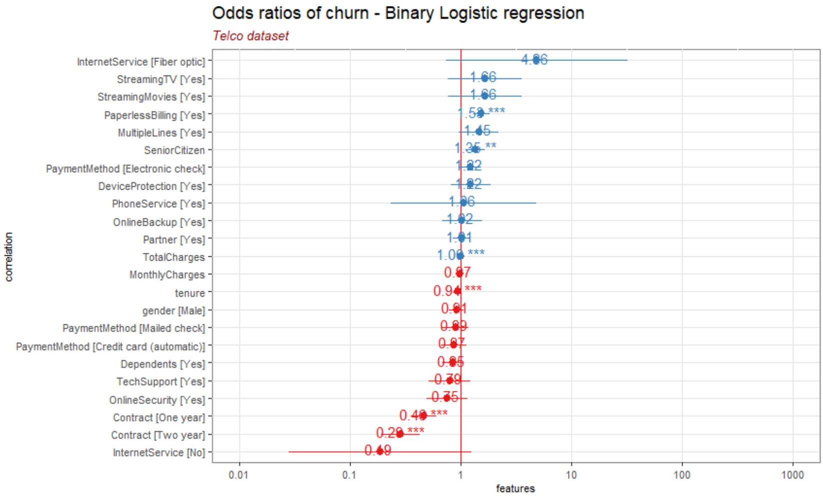
Modeling:
Logistic regression (2/5)

```r
# split dataset into training and testing sets
set.seed(2023)
ind <- sample(2, nrow(dataset), replace=TRUE, prob=c(0.7,0.3))
training <- dataset[ind==1,]
testing <- dataset[ind==2,]

# Binary logistic regression modeling - full model
model.1.lr <- glm(formula = ChurnDummy  ~ gender + SeniorCitizen + Partner +
                                          Dependents +
                                          tenure + PhoneService +  MultipleLines
      +
                                          InternetService +OnlineSecurity +
                                          OnlineBackup +
                                          DeviceProtection + TechSupport +
                                          StreamingTV +
                                          StreamingMovies + Contract +
                                          PaperlessBilling +
                                          PaymentMethod + MonthlyCharges +
                                          TotalCharges,
                  data = training, family = "binomial")

summary(model.1.lr)

# export the results in LaTex document
print(xtable(summary$coefficients, type = "latex"), file = "Customer_churn_
     analysis_tables.tex")
```

Modeling: Logistic regression (3/5)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.12	0.99	1.13	0.26
genderMale	-0.09	0.08	-1.22	0.22
SeniorCitizen	0.30	0.10	2.94	0.00
PartnerYes	0.01	0.09	0.14	0.89
DependentsYes	-0.17	0.11	-1.53	0.13
tenure	-0.06	0.01	-8.54	0.00
PhoneServiceYes	0.06	0.78	0.07	0.94
MultipleLinesYes	0.37	0.21	1.75	0.08
InternetServiceFiber optic	1.58	0.96	1.65	0.10
InternetServiceNo	-1.69	0.97	-1.74	0.08
OnlineSecurityYes	-0.29	0.22	-1.33	0.18
OnlineBackupYes	0.02	0.21	0.10	0.92
DeviceProtectionYes	0.20	0.21	0.95	0.34
TechSupportYes	-0.23	0.22	-1.08	0.28
StreamingTVYes	0.51	0.39	1.30	0.19
StreamingMoviesYes	0.51	0.39	1.29	0.20
ContractOne year	-0.77	0.13	-5.89	0.00
ContractTwo year	-1.25	0.21	-6.08	0.00
PaperlessBillingYes	0.43	0.09	4.82	0.00
PaymentMethodCredit card (automatic)	-0.14	0.14	-1.04	0.30
PaymentMethodElectronic check	0.20	0.11	1.78	0.08
PaymentMethodMailed check	-0.11	0.14	-0.81	0.42
MonthlyCharges	-0.03	0.04	-0.89	0.37
TotalCharges	0.00	0.00	4.06	0.00

Modeling: Logistic regression (4/5)

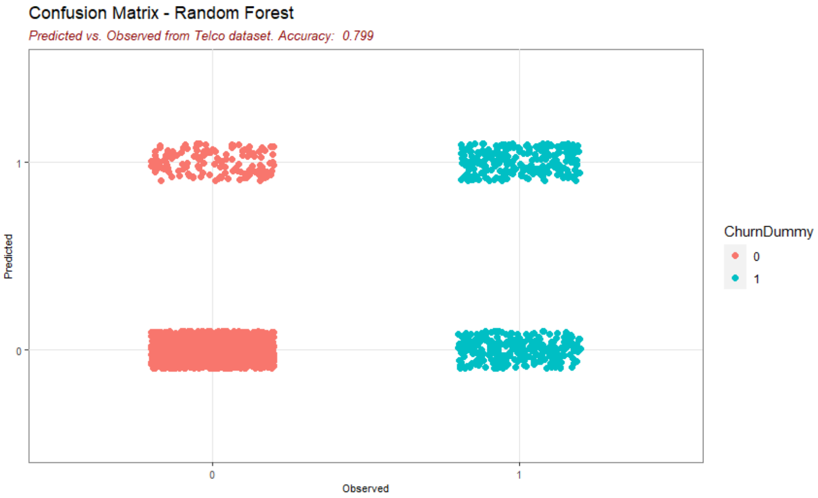


Modeling: Logistic regression (4/5)

Confusion matrix: accuracy = 0.815

	0	1
0	1365	237
1	140	293

Modeling: Random forest (1/3)



Modeling: Random forest (2/3)

Confusion matrix: accuracy $= 0.79$

	0	1
0	1373	277
1	132	253

Modeling: Logistic regression (1/4)

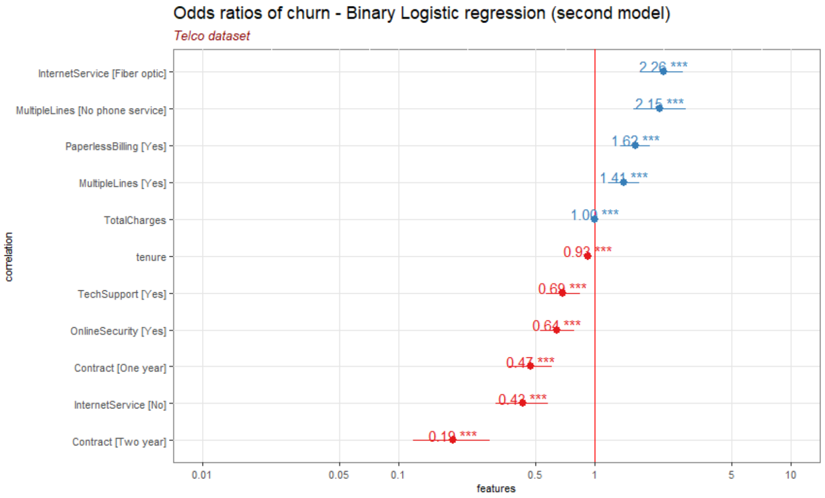
Some variables may not be relevant to the model or have low explanatory power. **Stepwise model selection** provides one possible solution to select our covariates based on Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) reduction (not available for Quasipoisson models).

```
library(MASS)
model.2.lr <- stepAIC(model.1.lr, direction = 'both',
                               k = log(dim(training)[1]))

                   Df Deviance    AIC
<none>                  4139.6 4241.8
+ PaymentMethod     3   4114.4 4242.2
+ SeniorCitizen     1   4133.2 4243.9
+ OnlineBackup      1   4133.3 4244.1
+ Dependents        1   4133.4 4244.1
+ StreamingMovies   1   4133.6 4244.3
+ StreamingTV       1   4134.3 4245.0
- TechSupport       1   4153.3 4247.0
+ MonthlyCharges    1   4136.4 4247.1
+ Partner           1   4137.7 4248.5
+ DeviceProtection  1   4139.3 4250.1
+ gender            1   4139.5 4250.2
- OnlineSecurity    1   4159.2 4252.9
- MultipleLines     2   4171.8 4257.0
- PaperlessBilling  1   4170.0 4263.7
- TotalCharges      1   4178.1 4271.7
- InternetService   1   4191.5 4285.1
- Contract          2   4220.4 4305.6
- tenure            1   4265.2 4358.9
```

Modeling: Logistic regression (2/4)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.38	0.12	-3.27	0.00
tenure	-0.08	0.01	-10.06	0.00
MultipleLinesNo phone service	0.76	0.16	4.81	0.00
MultipleLinesYes	0.34	0.09	3.66	0.00
InternetServiceFiber optic	0.81	0.11	7.14	0.00
InternetServiceNo	-0.85	0.16	-5.42	0.00
OnlineSecurityYes	-0.44	0.10	-4.39	0.00
TechSupportYes	-0.37	0.10	-3.68	0.00
ContractOne year	-0.75	0.13	-5.83	0.00
ContractTwo year	-1.66	0.23	-7.34	0.00
PaperlessBillingYes	0.48	0.09	5.48	0.00
TotalCharges	0.00	0.00	5.95	0.00

Modeling: Logistic regression (3/4)



Modeling: Logistic regression (4/4)

Confusion matrix: accuracy $= 0.814$

	0	1
0	1362	235
1	143	295

Conclusions

(i) Customers who have a short term subscription appear to churn much more.

(ii) The best model in terms of accuracy is the first binary logistic regression model, the full model. It can be used for prediction

(iii) To be continued

References

https://rpubs.com/anitaowens/customerchurn

https://colorado.posit.co/rsc/churn/modeling/tensorflow-w-r.nb.html

The R Project for Statistical Computing:
https://www.r-project.org/