Regression Analysis
Regression Analysis in Practice

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Predicting Churn Values of Customers: Regression & Variable Selection

Georgia Tech

About This Lesson

Georgia Tech

Logistic Regression

Create full model
full.model <- glm(Churn.Value~ ., family = "binomial", data = train)
summary(full.model)
Finding insignificant variables
which(summary(full.model)$coeff[,4]>0.05)

The overall regression seems to have explanatory power
Model Assessment: Multicollinearity
vifs <- vif(full.model)

Not statistically significant in the full model:
Gender, Senior Citizen, Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Payment Method, Monthly Charges

Logistic Regression (cont'd)

Create full model
full.model <- glm(Churn.Value~ ., family = "binomial", data = train)
summary(full.model)
Finding insignificant variables
which(summary(full.model)$coeff[,4]>0.05)

The overall regression seems to have explanatory power
Model Assessment: Multicollinearity
vifs <- vif(full.model)

	GVIF	Df	GVIF^(1/(2*Df))
Gender	1.003414	1	1.001705
`Senior Citizen`	1.112401	1	1.054704
Partner	1.248636	1	1.117424
Dependents	1.098666	1	1.048173
`Tenure Months`	15.612548	1	3.951272
`Phone Service`	35.526189	1	5.960385
`Multiple Lines`	7.434935	1	2.726708
`Internet Service`	382.924211	2	4.423624
`Online Security`	5.158636	1	2.271263
`Online Backup`	6.520493	1	2.553526
`Device Protection`	6.611606	1	2.571304
`Tech Support`	5.409603	1	2.325855
`Streaming TV`	25.075402	1	5.007534
`Streaming Movies`	25.317771	1	5.031677
Contract	1.625406	2	1.129121
`Paperless Billing`	1.128532	1	1.062324
`Payment Method`	1.413278	3	1.059346
`Monthly Charges`	694.903171	1	26.361016
`Total Charges`	20.166529	1	4.490716

Variable Selection

Reduce the number of factors in the model

1. Overfitting
   - Model with large # of factors can fit too closely, cause random effects
   - It can cause bad estimates
2. Simplicity
   - Less chance of insignificant factors
   - Easier to interpret

Variable Selection (cont'd)

- Forward-Backward Stepwise Regression

```
# Create minimum model including an intercept
min.model <-  glm(Churn.Value~ 1, family = "binomial", data = train)
# Perform stepwise regression
step.model <- step(min.model, scope = list(lower = min.model, upper = full.model),
            direction = "both", trace = FALSE)
```

- **Not selected**: Gender, Senior Citizen, Online Backup, Device Protection, Monthly Charges
- **Not statistically significant**: Payment Method by Mailed check and by Credit Card

Variable Selection (cont'd)

- LASSO Regression

```
# Set predictors and response to correct format
x.train <- model.matrix(Churn.Value ~ ., train)[,-1]
y.train <- train$Churn.Value
# Use cross validation to find optimal lambda
cv.lasso <- cv.glmnet(x.train, y.train, alpha = 1, family = "binomial")
# Train Lasso and display coefficients with optimal lambda
lasso.model <- glmnet(x.train, y.train, alpha = 1, family = "binomial")
coef(lasso.model, cv.lasso$lambda.min)
```

- Elastic Net Regression

```
# Use cross validation to find optimal lambda
cv.elnet <- cv.glmnet(x.train, y.train, alpha = 0.5, family = "binomial")
# Train Elastic Net and display coefficients with optimal lambda
elnet.model <- glmnet(x.train, y.train, alpha = 0.5, family = "binomial")
coef(elnet.model, cv.elnet$lambda.min)
```

- **Not selected for both models**: Monthly Charges

Summary