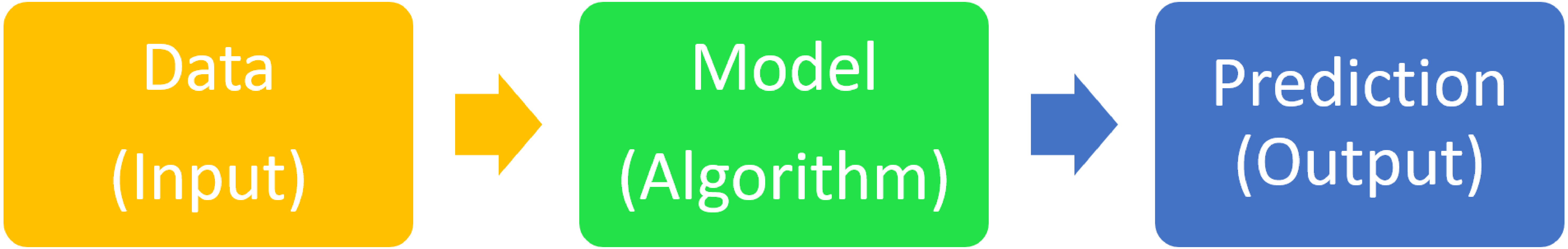
Training and testing datasets: splitting data

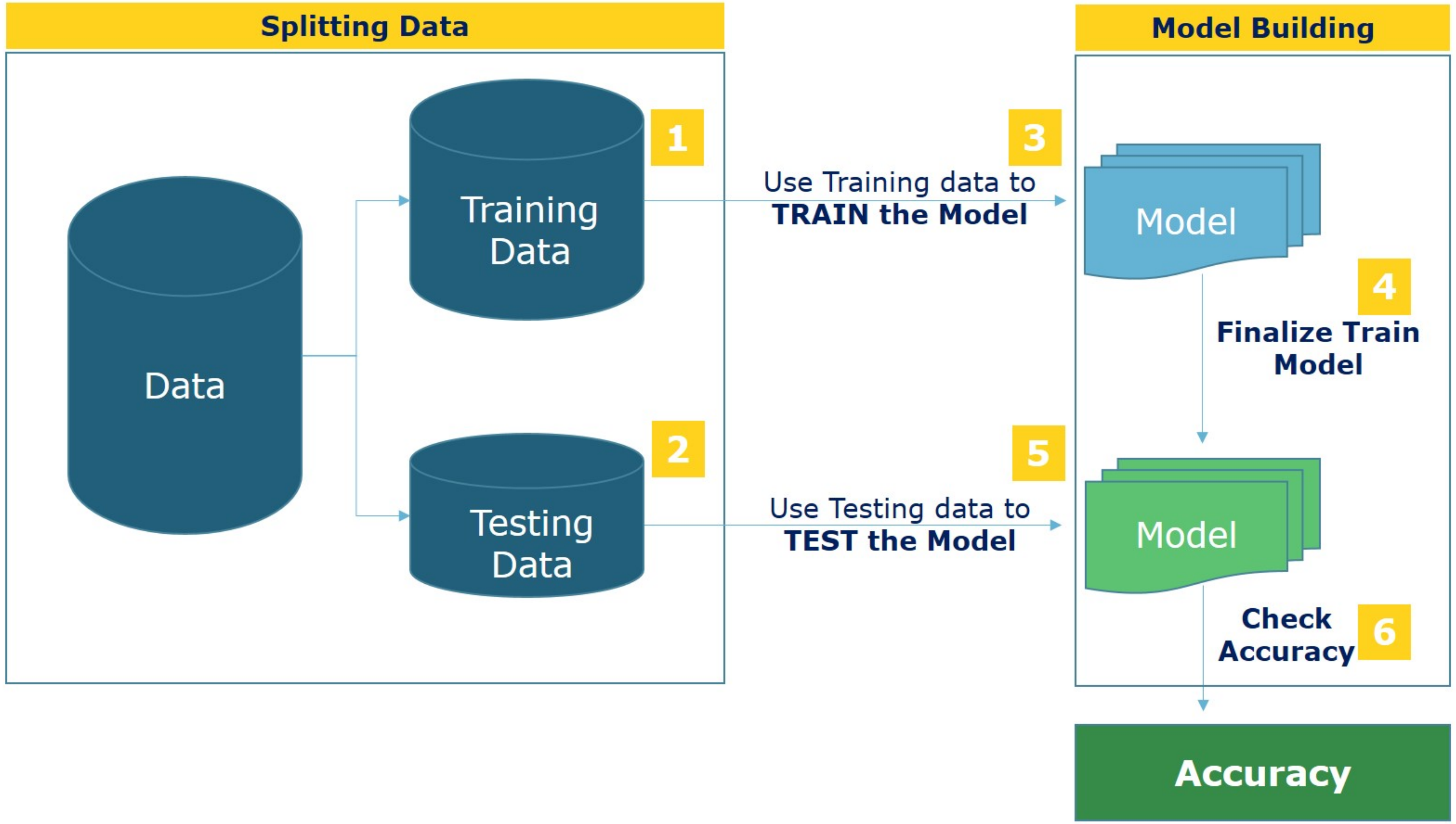
HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R

Anurag Gupta
People Analytics Practitioner

What is a model?



Why Split Data into Train & Test sets ?



Splitting data with caret

```
# Load caret
library(caret)

# Set seed
set.seed(567)

# Store row numbers for training dataset
index_train <- createDataPartition(emp_final$turnover, p = 0.5, list = FALSE)

# Create training dataset
train_set <- emp_final[index_train, ]

# Create testing dataset
test_set <- emp_final[-index_train, ]
```

Let's practice!

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R

Introduction to logistic regression

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R

Anurag Gupta
People Analytics Practitioner

What is logistic regression?

- Classification technique
- Predicts the probability of occurrence of an event
- Dependent variable is categorical

Understanding logistic regression

- Independent variables
  - Continuous / Categorical
  - age, tenure, compensation, level etc.
- Dependent variable
  - Binary / Dichotomous variable
  - turnover (1, 0)

Building a simple logistic regression model

```
simple_log <- glm(turnover ~ emp_age,
                  family = "binomial", data = train_set)
```

```
summary(simple_log)
```

```
Call:
glm(formula = turnover ~ emp_age, family = "binomial", data = train_set)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9431  -0.7406  -0.6107  -0.4006   2.4334

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.58131    0.58684   4.399 1.09e-05 ***
emp_age     -0.13864    0.02093  -6.623 3.52e-11 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1389.4  on 1367  degrees of freedom
Residual deviance: 1338.6  on 1366  degrees of freedom
AIC: 1342.6

Number of Fisher Scoring iterations: 4
```

Removing variables

- `emp_id` , `mgr_id` (ID columns)
- `date_of_joining` , `last_working_date` , `cutoff_date` (`tenure` is a linear combination of these columns)
- `median_compensation` (directly related to `level`)
- `mgr_age` , `emp_age` (`age_diff` is a linear combination of these columns)
- `department` (only one possible value)
- `status` (same as `turnover`)

Removing variables

```r
# Drop variables and save the resulting object as train_set_multi
train_set_multi <- train_set %>%
  select(-c(emp_id, mgr_id,
            date_of_joining, last_working_date, cutoff_date,
            mgr_age, emp_age,
            median_compensation,
            department, status))
```

Building multiple logistic regression model

```
multi_log <- glm(turnover ~ ., family = "binomial",
                 data = train_set_multi)
```

```r
summary(multi_log)
```

```
Call:
glm(formula = turnover ~ ., family = "binomial", data = train_set_multi)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4235  -0.1392  -0.0345  -0.0001   3.4580
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -1.348e+01  4.813e+00  -2.800 0.005104 **
locationNew York                1.264e+00  4.655e-01   2.715 0.006624 **
locationOrlando                -1.031e+00  4.200e-01  -2.455 0.014077 *
levelSpecialist                 1.583e+01  9.695e+02   0.016 0.986971
percent_hike                   -5.669e-01  8.102e-02  -6.997 2.61e-12 ***
tenure                         -5.863e-01  1.192e-01  -4.920 8.65e-07 ***
total_experience                8.598e-02  8.380e-02   1.026 0.304871
.....
# We removed several variables for brevity
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
 Null deviance: 1389.37  on 1367  degrees of freedom
Residual deviance:  326.66  on 1326  degrees of freedom
AIC: 410.66

Number of Fisher Scoring iterations: 18
```

Let's practice!

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R

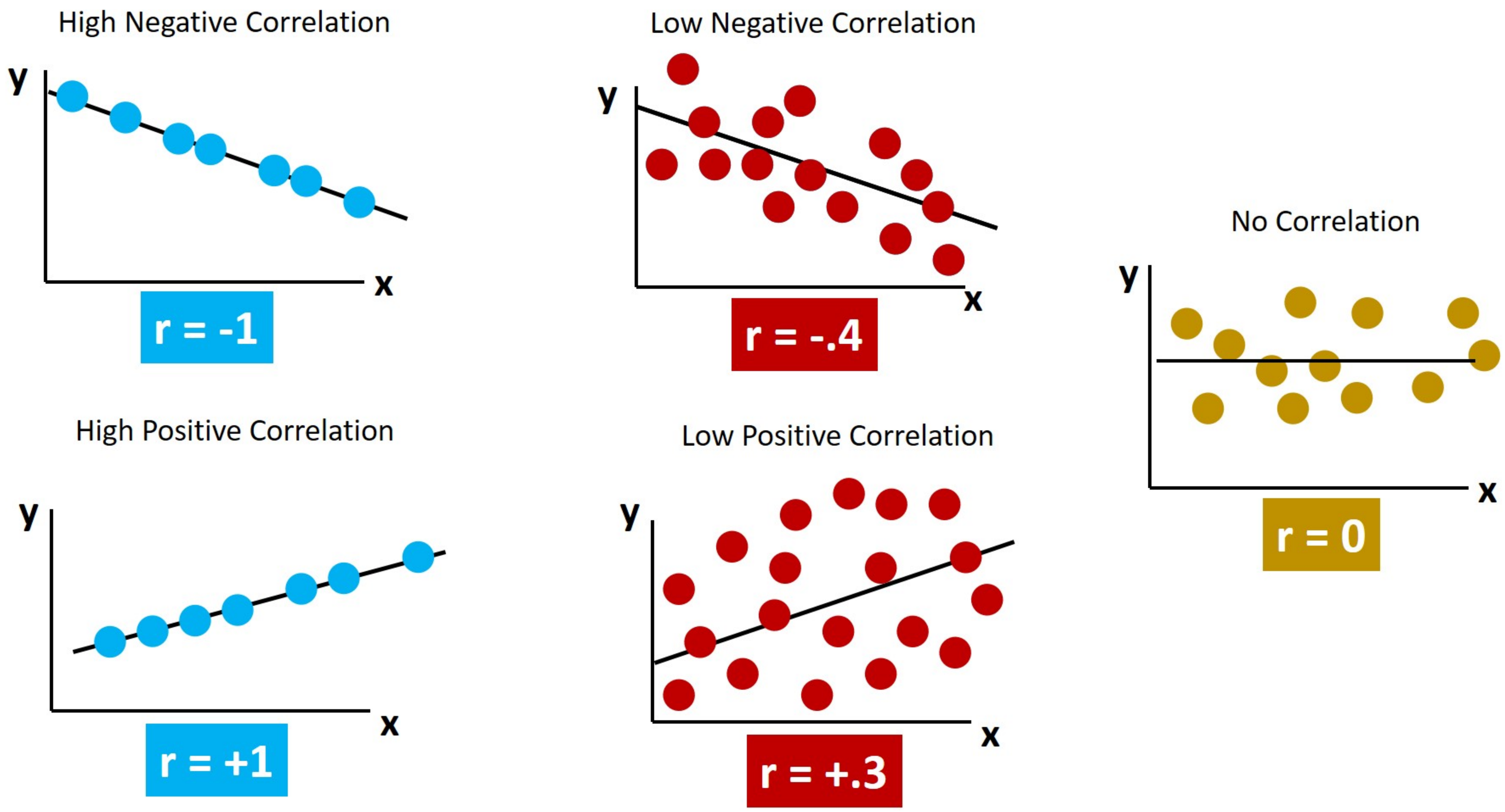
Detecting and dealing with multicollinearity

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R

Anurag Gupta
People Analytics Practitioner

Understanding correlation

> Correlation is the measure of association between two numeric variables


High Negative Correlation
y
x
r = -1
Low Negative Correlation
y
x
r = -.4
No Correlation
y
x
r = 0
High Positive Correlation
y
x
r = +1
Low Positive Correlation
y
x
r = +.3

Calculating correlation in R

```
# Calculate the correlation coefficient
cor(train_set$emp_age, train_set$compensation)
```

```
0.6117855
```

What is multicollinearity?

> Multicollinearity occurs when one independent variable is highly collinear with a set of two or more independent variables.

How to detect multicollinearity?

VIF (VARIANCE INFLATION FACTOR)

```
# Load car package
library(car)

# Logistic regression model
multi_log <- glm(turnover ~ ., family = "binomial",
                 data = train_set_multi)

# Calculate VIF
vif(multi_log)
```

Variance inflation factor

```
                                    GVIF Df GVIF^(1/(2*Df))
location                    2.318640e+00  2        1.233981
level                       5.716850e+06  1     2390.993458
gender                      1.262625e+00  1        1.123666
rating                      4.381767e+00  4        1.202835
mgr_rating                  2.471489e+00  4        1.119747
mgr_reportees               1.314709e+00  1        1.146608
mgr_tenure                  1.278559e+00  1        1.130734
compensation                3.998338e+01  1        6.323241
percent_hike                3.167576e+00  1        1.779769
hiring_score                1.143613e+00  1        1.069399
hiring_source               2.000099e+00  6        1.059467
no_previous_companies_worked 3.291703e+00  1        1.814305
distance_from_home          1.355795e+00  1        1.164386
total_dependents            1.930188e+00  1        1.389312
marital_status              2.320518e+00  1        1.523325
education                   1.460697e+00  1        1.208593
.....
```

Rule of thumb for interpreting VIF value

VIF	Interpretation
1	Not correlated
Between 1 and 5	Moderately correlated
Greater than 5	Highly correlated

How to deal with multicollinearity?

- Step 1: Calculate VIF of the model
- Step 2: Identify if any variable has VIF greater than 5
    - Step 2a: Remove the variable from the model if it has a VIF of 5
    - Step 2b: If there are multiple variables with VIF greater than 5, only remove the variable with the highest VIF
- Step 3: Repeat steps 1 and 2 until VIF of each variable is less than 5

Removing a variable from a model

```
new_model <- glm(dependent_variable ~ . - variable_to_remove,
                 family = "binomial", data = dataset)
```

Let's practice!

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R

Final steps to nirvana

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R

Anurag Gupta
People Analytics Practitioner

Build your final model

```
# Final model, you will complete this in the next exercise
final_log <- glm(...)
```

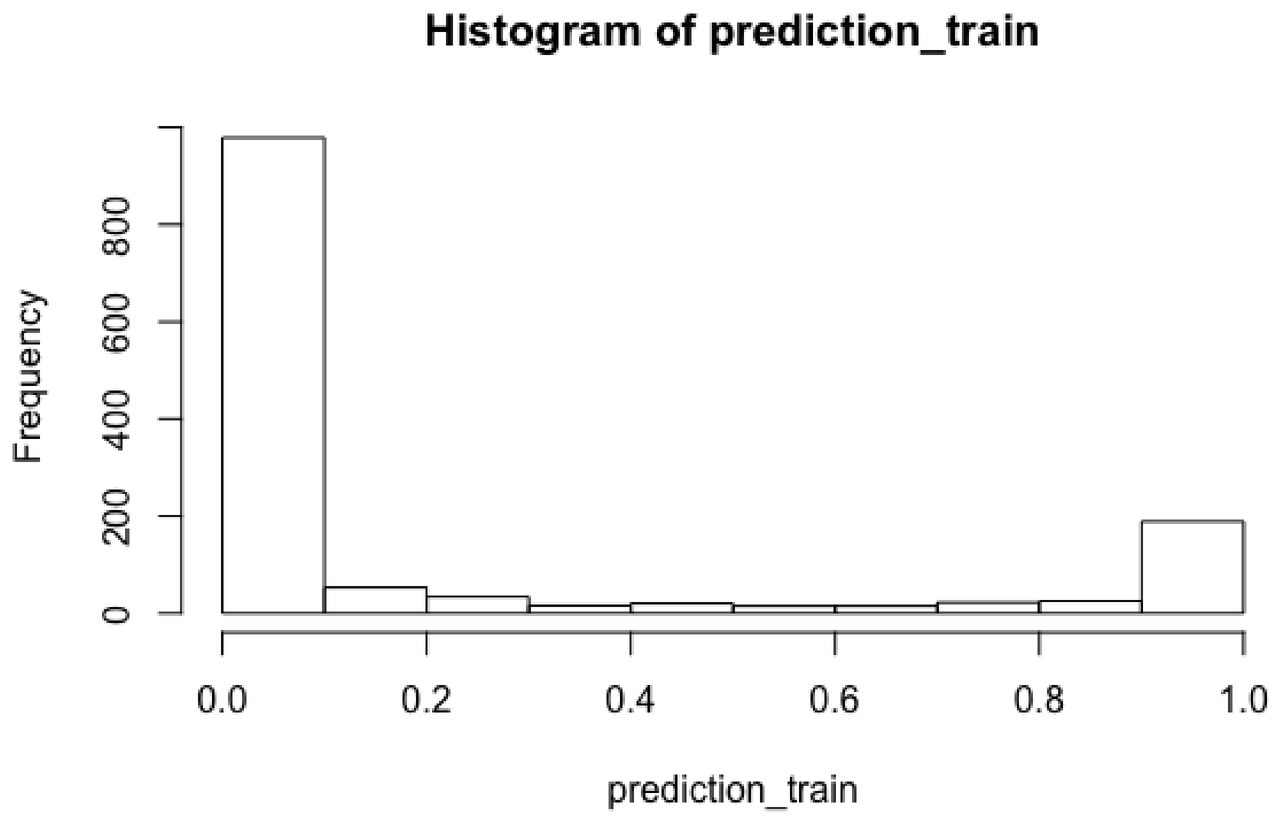
Predicting probability of turnover

```r
# Make predictions for training dataset
prediction_train <- predict(final_log, newdata = train_set_final,
                            type = "response")
prediction_train[c(205, 645)]
```

```
       205        645
0.06069079 0.99999898
```

Plot probability range: training dataset

```
# Look at the predictions range
hist(prediction_train)
```

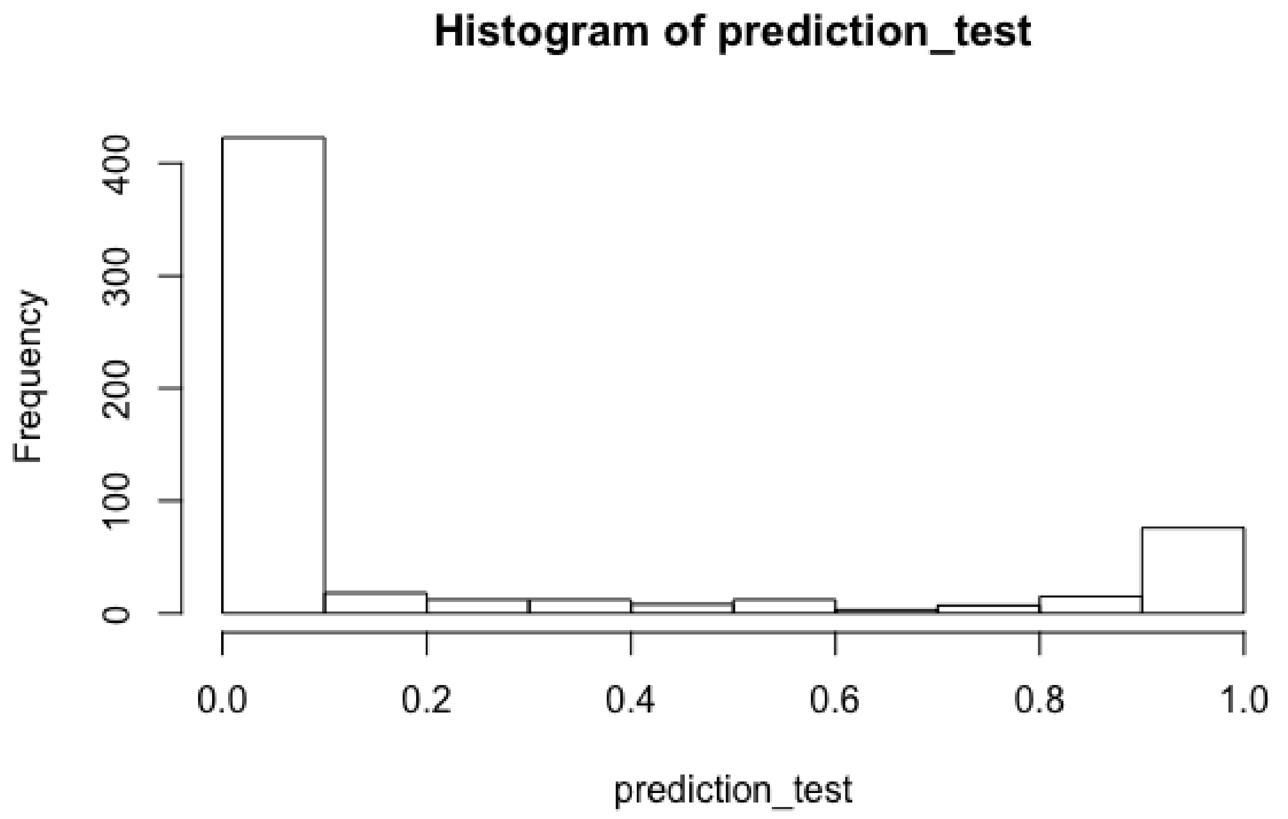


Predicting probability: testing dataset

```
# Make predictions for testing dataset
# test_set is the test dataset from chapter 3 exercise 2
prediction_test <- predict(final_log, newdata = test_set,
                           type = "response")
```

Plot probability range: testing dataset

```
# Look at the predictions range
hist(prediction_test)
```



Let's practice!

HUMAN RESOURCES ANALYTICS: PREDICTING EMPLOYEE CHURN IN R