R Notebook

require(tidymodels)

## Loading required package: tidymodels

## ── Attaching packages ────────────────────────────────────── tidymodels 1.0.0 ──

## ✔ broom 1.0.2 ✔ recipes 1.0.4   
## ✔ dials 1.1.0 ✔ rsample 1.1.1   
## ✔ dplyr 1.0.10 ✔ tibble 3.1.8   
## ✔ ggplot2 3.4.1 ✔ tidyr 1.2.1   
## ✔ infer 1.0.4 ✔ tune 1.0.1   
## ✔ modeldata 1.1.0 ✔ workflows 1.1.2   
## ✔ parsnip 1.0.3 ✔ workflowsets 1.0.0   
## ✔ purrr 0.3.5 ✔ yardstick 1.1.0

## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## ✖ purrr::discard() masks scales::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ recipes::step() masks stats::step()  
## • Use tidymodels\_prefer() to resolve common conflicts.

require(tidyverse)

## Loading required package: tidyverse

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ readr 2.1.3 ✔ forcats 0.5.2  
## ✔ stringr 1.5.0   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ readr::col\_factor() masks scales::col\_factor()  
## ✖ purrr::discard() masks scales::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ stringr::fixed() masks recipes::fixed()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ readr::spec() masks yardstick::spec()

library(tidyverse)  
#install.packages("hablar")  
#library(hablar)  
#install.packages("skimr")  
#library(skimr)  
#require(tidyverse)  
#library(tidyverse)  
#library(gapminder)#  
#install.packages("hablar")  
#library(hablar)  
#library(dplyr)  
#remove.packages(c("ggplot2", "data.table"))  
#install.packages('Rcpp', dependencies = TRUE)  
#install.packages('ggplot2', dependencies = TRUE)  
#install.packages('data.table', dependencies = TRUE)  
library(ggplot2)  
#library(dplyr)  
#install.packages("zoo")  
library(vip)

##   
## Attaching package: 'vip'  
##   
## The following object is masked from 'package:utils':  
##   
## vi

#install.packages("writexl")  
#library(writexl)  
#install.packages("see")

## Introduction to tidy models

Bank\_churners<-read\_csv("C:/Users/Lenovo/Downloads/BankChurners.csv")

## Rows: 5998 Columns: 21  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (6): Attrition\_Flag, Gender, Education\_Level, Marital\_Status, Income\_Ca...  
## dbl (15): CLIENTNUM, Customer\_Age, Dependent\_count, Months\_on\_book, Total\_Re...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

glimpse(Bank\_churners)

## Rows: 5,998  
## Columns: 21  
## $ CLIENTNUM <dbl> 768805383, 818770008, 713982108, 769911858,…  
## $ Attrition\_Flag <chr> "Existing Customer", "Existing Customer", "…  
## $ Customer\_Age <dbl> 45, 49, 51, 40, 40, 44, 51, 32, 37, 48, 42,…  
## $ Gender <chr> "M", "F", "M", "F", "M", "M", "M", "M", "M"…  
## $ Dependent\_count <dbl> 3, 5, 3, 4, 3, 2, 4, 0, 3, 2, 5, 1, 1, 3, 2…  
## $ Education\_Level <chr> "High School", "Graduate", "Graduate", "Hig…  
## $ Marital\_Status <chr> "Married", "Single", "Married", "Unknown", …  
## $ Income\_Category <chr> "$60K - $80K", "Less than $40K", "$80K - $1…  
## $ Card\_Category <chr> "Blue", "Blue", "Blue", "Blue", "Blue", "Bl…  
## $ Months\_on\_book <dbl> 39, 44, 36, 34, 21, 36, 46, 27, 36, 36, 31,…  
## $ `Total\_Relationship Count` <dbl> 5, 6, 4, 3, 5, 3, 6, 2, 5, 6, 5, 6, 3, 5, 5…  
## $ Months\_Inactive\_12\_mon <dbl> 1, 1, 1, 4, 1, 1, 1, 2, 2, 3, 3, 2, 6, 1, 2…  
## $ Contacts\_Count\_12\_mon <dbl> 3, 2, 0, 1, 0, 2, 3, 2, 0, 3, 2, 3, 0, 3, 2…  
## $ Credit\_Limit <dbl> 12691.0, 8256.0, 3418.0, 3313.0, 4716.0, 40…  
## $ Total\_Revolving\_Bal <dbl> 777, 864, 0, 2517, 0, 1247, 2264, 1396, 251…  
## $ Avg\_Open\_To\_Buy <dbl> 11914.0, 7392.0, 3418.0, 796.0, 4716.0, 276…  
## $ Total\_Amt\_Chng\_Q4\_Q1 <dbl> 1.335, 1.541, 2.594, 1.405, 2.175, 1.376, 1…  
## $ Total\_Trans\_Amt <dbl> 1144, 1291, 1887, 1171, 816, 1088, 1330, 15…  
## $ Total\_Trans\_Ct <dbl> 42, 33, 20, 20, 28, 24, 31, 36, 24, 32, 42,…  
## $ Total\_Ct\_Chng\_Q4\_Q1 <dbl> 1.625, 3.714, 2.333, 2.333, 2.500, 0.846, 0…  
## $ Avg\_Utilization\_Ratio <dbl> 0.061, 0.105, 0.000, 0.760, 0.000, 0.311, 0…

Bank\_churners%>%head()

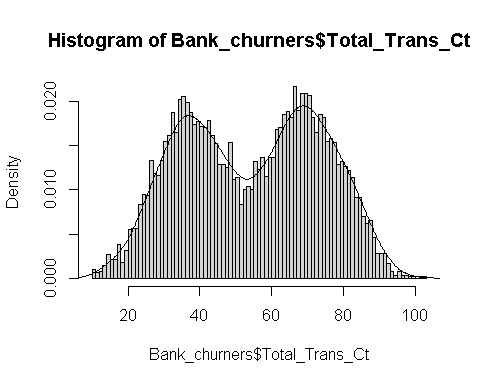
## # A tibble: 6 × 21  
## CLIEN…¹ Attri…² Custo…³ Gender Depen…⁴ Educa…⁵ Marit…⁶ Incom…⁷ Card\_…⁸ Month…⁹  
## <dbl> <chr> <dbl> <chr> <dbl> <chr> <chr> <chr> <chr> <dbl>  
## 1 7.69e8 Existi… 45 M 3 High S… Married $60K -… Blue 39  
## 2 8.19e8 Existi… 49 F 5 Gradua… Single Less t… Blue 44  
## 3 7.14e8 Existi… 51 M 3 Gradua… Married $80K -… Blue 36  
## 4 7.70e8 Existi… 40 F 4 High S… Unknown Less t… Blue 34  
## 5 7.09e8 Existi… 40 M 3 Uneduc… Married $60K -… Blue 21  
## 6 7.13e8 Existi… 44 M 2 Gradua… Married $40K -… Blue 36  
## # … with 11 more variables: `Total\_Relationship Count` <dbl>,  
## # Months\_Inactive\_12\_mon <dbl>, Contacts\_Count\_12\_mon <dbl>,  
## # Credit\_Limit <dbl>, Total\_Revolving\_Bal <dbl>, Avg\_Open\_To\_Buy <dbl>,  
## # Total\_Amt\_Chng\_Q4\_Q1 <dbl>, Total\_Trans\_Amt <dbl>, Total\_Trans\_Ct <dbl>,  
## # Total\_Ct\_Chng\_Q4\_Q1 <dbl>, Avg\_Utilization\_Ratio <dbl>, and abbreviated  
## # variable names ¹​CLIENTNUM, ²​Attrition\_Flag, ³​Customer\_Age,  
## # ⁴​Dependent\_count, ⁵​Education\_Level, ⁶​Marital\_Status, ⁷​Income\_Category, …

summary(Bank\_churners)

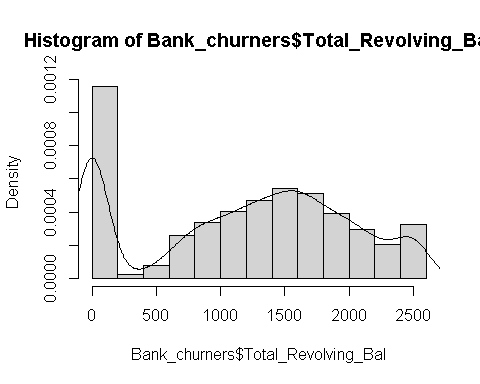
## CLIENTNUM Attrition\_Flag Customer\_Age Gender   
## Min. :708083283 Length:5998 Min. :26.00 Length:5998   
## 1st Qu.:713095796 Class :character 1st Qu.:40.00 Class :character   
## Median :718055433 Mode :character Median :46.00 Mode :character   
## Mean :739795002 Mean :46.38   
## 3rd Qu.:778343152 3rd Qu.:52.00   
## Max. :828343083 Max. :73.00   
## Dependent\_count Education\_Level Marital\_Status Income\_Category   
## Min. :0.000 Length:5998 Length:5998 Length:5998   
## 1st Qu.:1.000 Class :character Class :character Class :character   
## Median :2.000 Mode :character Mode :character Mode :character   
## Mean :2.278   
## 3rd Qu.:3.000   
## Max. :5.000   
## Card\_Category Months\_on\_book Total\_Relationship Count  
## Length:5998 Min. :13.00 Min. :1.00   
## Class :character 1st Qu.:31.00 1st Qu.:3.00   
## Mode :character Median :36.00 Median :4.00   
## Mean :35.98 Mean :4.31   
## 3rd Qu.:41.00 3rd Qu.:5.00   
## Max. :56.00 Max. :6.00   
## Months\_Inactive\_12\_mon Contacts\_Count\_12\_mon Credit\_Limit   
## Min. :0.000 Min. :0.000 Min. : 1438   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.: 2644   
## Median :2.000 Median :3.000 Median : 5013   
## Mean :2.299 Mean :2.608 Mean : 8714   
## 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.:11335   
## Max. :6.000 Max. :6.000 Max. :34516   
## Total\_Revolving\_Bal Avg\_Open\_To\_Buy Total\_Amt\_Chng\_Q4\_Q1 Total\_Trans\_Amt  
## Min. : 0.0 Min. : 3 Min. :0.0000 Min. : 510   
## 1st Qu.: 580.5 1st Qu.: 1368 1st Qu.:0.6130 1st Qu.:1711   
## Median :1300.5 Median : 3782 Median :0.7380 Median :2566   
## Mean :1183.5 Mean : 7530 Mean :0.7736 Mean :2816   
## 3rd Qu.:1787.8 3rd Qu.:10081 3rd Qu.:0.8860 3rd Qu.:3989   
## Max. :2517.0 Max. :34516 Max. :3.3970 Max. :5783   
## Total\_Trans\_Ct Total\_Ct\_Chng\_Q4\_Q1 Avg\_Utilization\_Ratio  
## Min. : 10.00 Min. :0.0000 Min. :0.000   
## 1st Qu.: 38.00 1st Qu.:0.5470 1st Qu.:0.035   
## Median : 56.00 Median :0.6820 Median :0.178   
## Mean : 54.79 Mean :0.7065 Mean :0.276   
## 3rd Qu.: 71.00 3rd Qu.:0.8210 3rd Qu.:0.496   
## Max. :103.00 Max. :3.7140 Max. :0.999

#skim(Bank\_churners)

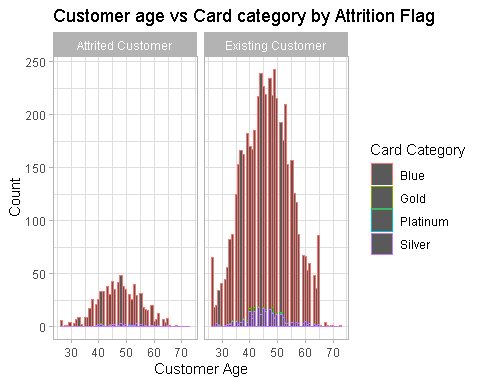
hist(Bank\_churners$Total\_Trans\_Ct,prob=TRUE,breaks=100)  
lines(density(Bank\_churners$Total\_Trans\_Ct))



hist(Bank\_churners$Total\_Revolving\_Bal,prob=TRUE,breaks=9)  
lines(density(Bank\_churners$Total\_Revolving\_Bal))



ggplot(data = Bank\_churners,  
 mapping = aes(x = Customer\_Age,color = Card\_Category)) +geom\_histogram(binwidth = 0.8) +facet\_wrap(~ Card\_Category) +labs(title =  
"Customer age vs Card category by Attrition Flag",x = "Customer Age",y = "Count") +  
 scale\_color\_discrete(name = "Card Category") +theme\_light() +facet\_wrap(~ Card\_Category) +facet\_grid(. ~ Attrition\_Flag)



Bank\_churners$Attrition\_Flag <- as.numeric(factor(Bank\_churners$Attrition\_Flag))  
  
set.seed(123)  
  
# Create a split object  
Bankchurners\_split <- initial\_split(Bank\_churners, prop = 0.90,   
 strata = Attrition\_Flag)  
  
# Build training data set  
Bankchurners\_training <- Bankchurners\_split %>%   
 training()  
  
# Build testing data set  
Bankchurners\_test <- Bankchurners\_split %>%   
 testing()

lm\_model <- linear\_reg() %>%   
 set\_engine('lm') %>% # adds lm implementation of linear regression  
 set\_mode('regression')

lm\_fit <- lm\_model %>%   
 fit(Attrition\_Flag ~ . , data=Bankchurners\_training)  
# View lm\_fit properties  
lm\_fit

## parsnip model object  
##   
##   
## Call:  
## stats::lm(formula = Attrition\_Flag ~ ., data = data)  
##   
## Coefficients:  
## (Intercept) CLIENTNUM   
## 9.460e-01 1.505e-10   
## Customer\_Age GenderM   
## -9.887e-04 7.362e-02   
## Dependent\_count Education\_LevelDoctorate   
## -1.398e-02 -5.171e-02   
## Education\_LevelGraduate Education\_LevelHigh School   
## -1.669e-03 -1.054e-03   
## Education\_LevelPost-Graduate Education\_LevelUneducated   
## -1.643e-02 -8.740e-03   
## Education\_LevelUnknown Marital\_StatusMarried   
## -1.321e-02 4.597e-02   
## Marital\_StatusSingle Marital\_StatusUnknown   
## -4.163e-03 4.899e-03   
## Income\_Category$40K - $60K Income\_Category$60K - $80K   
## 5.080e-02 2.464e-02   
## Income\_Category$80K - $120K Income\_CategoryLess than $40K   
## 2.298e-02 5.213e-02   
## Income\_CategoryUnknown Card\_CategoryGold   
## 3.922e-02 -2.427e-02   
## Card\_CategoryPlatinum Card\_CategorySilver   
## -1.090e-01 -1.465e-02   
## Months\_on\_book `Total\_Relationship Count`   
## 3.411e-04 6.234e-02   
## Months\_Inactive\_12\_mon Contacts\_Count\_12\_mon   
## -3.244e-02 -3.070e-02   
## Credit\_Limit Total\_Revolving\_Bal   
## 1.681e-06 8.390e-05   
## Avg\_Open\_To\_Buy Total\_Amt\_Chng\_Q4\_Q1   
## NA 4.902e-02   
## Total\_Trans\_Amt Total\_Trans\_Ct   
## -1.742e-05 7.073e-03   
## Total\_Ct\_Chng\_Q4\_Q1 Avg\_Utilization\_Ratio   
## 2.561e-01 1.110e-02

names(lm\_fit)

## [1] "lvl" "spec" "fit" "preproc" "elapsed"

To print a summary of our model, we can extract fit from lm\_fit and pass it to the summary() function. We can explore the estimated coefficients, F-statistics, p-values, residual standard error (also known as RMSE) and R2 value.

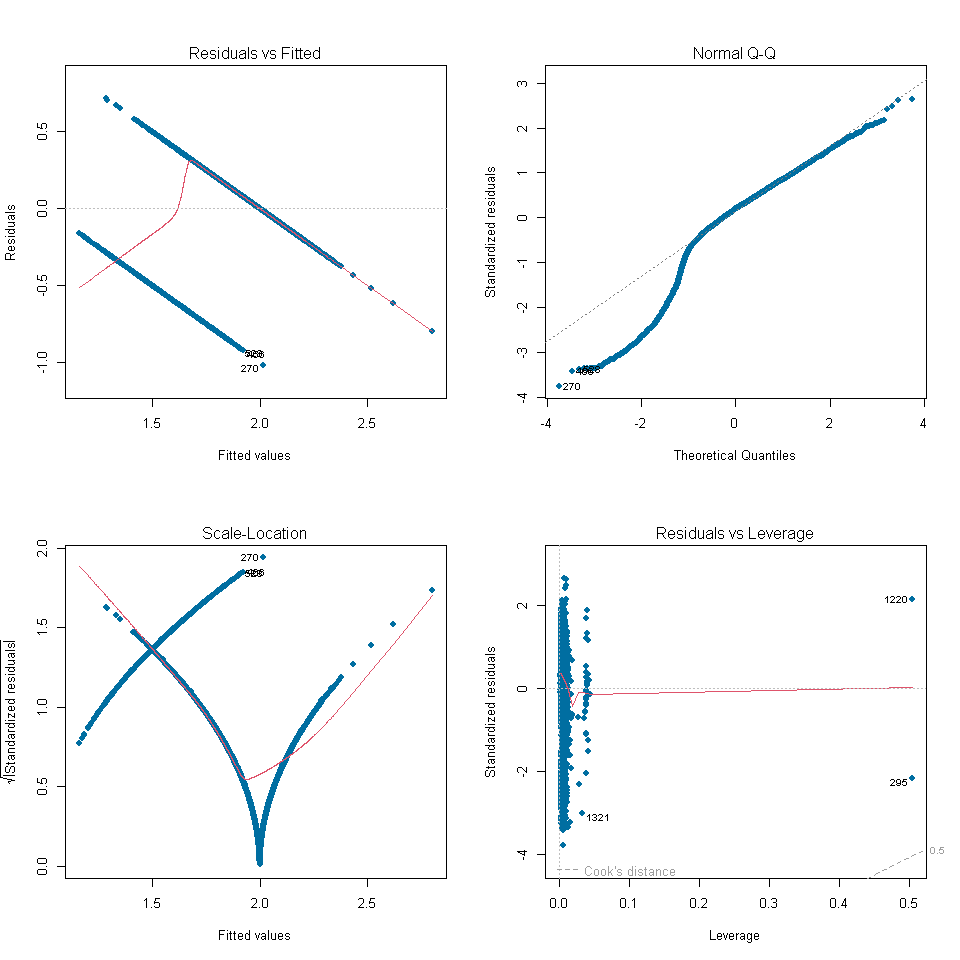
However, this feature is best for visually exploring our results on the training data since the results are returned as a data frame.

summary(lm\_fit$fit)

##   
## Call:  
## stats::lm(formula = Attrition\_Flag ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.01659 -0.09171 0.05278 0.17296 0.71362   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.460e-01 8.752e-02 10.808 < 2e-16 \*\*\*  
## CLIENTNUM 1.505e-10 1.012e-10 1.487 0.13715   
## Customer\_Age -9.887e-04 7.530e-04 -1.313 0.18923   
## GenderM 7.362e-02 1.344e-02 5.479 4.48e-08 \*\*\*  
## Dependent\_count -1.398e-02 2.918e-03 -4.793 1.69e-06 \*\*\*  
## Education\_LevelDoctorate -5.171e-02 2.098e-02 -2.464 0.01377 \*   
## Education\_LevelGraduate -1.669e-03 1.357e-02 -0.123 0.90209   
## Education\_LevelHigh School -1.054e-03 1.445e-02 -0.073 0.94183   
## Education\_LevelPost-Graduate -1.643e-02 2.051e-02 -0.801 0.42333   
## Education\_LevelUneducated -8.740e-03 1.521e-02 -0.575 0.56557   
## Education\_LevelUnknown -1.321e-02 1.521e-02 -0.869 0.38515   
## Marital\_StatusMarried 4.597e-02 1.471e-02 3.124 0.00179 \*\*   
## Marital\_StatusSingle -4.163e-03 1.482e-02 -0.281 0.77880   
## Marital\_StatusUnknown 4.899e-03 1.981e-02 0.247 0.80467   
## Income\_Category$40K - $60K 5.080e-02 1.891e-02 2.686 0.00725 \*\*   
## Income\_Category$60K - $80K 2.464e-02 1.703e-02 1.446 0.14814   
## Income\_Category$80K - $120K 2.298e-02 1.619e-02 1.419 0.15599   
## Income\_CategoryLess than $40K 5.213e-02 2.082e-02 2.503 0.01234 \*   
## Income\_CategoryUnknown 3.922e-02 2.208e-02 1.776 0.07581 .   
## Card\_CategoryGold -2.427e-02 5.213e-02 -0.466 0.64149   
## Card\_CategoryPlatinum -1.090e-01 1.926e-01 -0.566 0.57142   
## Card\_CategorySilver -1.465e-02 2.046e-02 -0.716 0.47383   
## Months\_on\_book 3.411e-04 7.605e-04 0.449 0.65375   
## `Total\_Relationship Count` 6.234e-02 2.968e-03 21.002 < 2e-16 \*\*\*  
## Months\_Inactive\_12\_mon -3.244e-02 3.734e-03 -8.686 < 2e-16 \*\*\*  
## Contacts\_Count\_12\_mon -3.070e-02 3.250e-03 -9.446 < 2e-16 \*\*\*  
## Credit\_Limit 1.681e-06 6.890e-07 2.440 0.01474 \*   
## Total\_Revolving\_Bal 8.390e-05 6.811e-06 12.318 < 2e-16 \*\*\*  
## Avg\_Open\_To\_Buy NA NA NA NA   
## Total\_Amt\_Chng\_Q4\_Q1 4.902e-02 1.624e-02 3.018 0.00256 \*\*   
## Total\_Trans\_Amt -1.742e-05 6.595e-06 -2.642 0.00826 \*\*   
## Total\_Trans\_Ct 7.073e-03 4.221e-04 16.756 < 2e-16 \*\*\*  
## Total\_Ct\_Chng\_Q4\_Q1 2.561e-01 1.512e-02 16.935 < 2e-16 \*\*\*  
## Avg\_Utilization\_Ratio 1.110e-02 2.319e-02 0.479 0.63220   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2709 on 5364 degrees of freedom  
## Multiple R-squared: 0.3594, Adjusted R-squared: 0.3556   
## F-statistic: 94.06 on 32 and 5364 DF, p-value: < 2.2e-16

We can use the plot() function to obtain diagnostic plots for our trained regression model. Again, we must first extract the fit object from lm\_fit and then pass it into plot(). These plots provide a check for the main assumptions of the linear regression model.

par(mfrow=c(2,2)) # plot all 4 plots in one  
  
plot(lm\_fit$fit,   
 pch = 16, # optional parameters to make points blue  
 col = '#006EA1')



### Tidy Training Results

To obtain the detailed results from our trained linear regression model in a data frame, we can use the tidy() and glance() functions directly on our trained parsnip model, lm\_fit.

The tidy() function takes a linear regression object and returns a data frame of the estimated model coefficients and their associated F-statistics and p-values.

The glance() function will return performance metrics obtained on the training data such as the R2 value (r.squared) and the RMSE (sigma).

# Data frame of estimated coefficients  
tidy(lm\_fit)

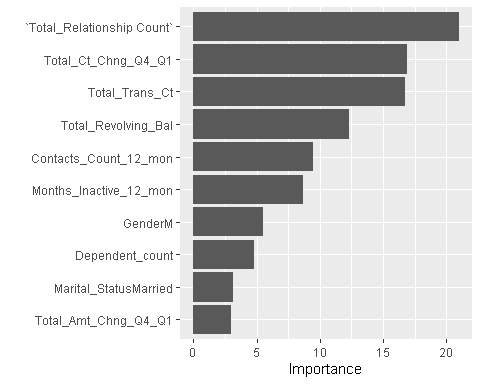
## # A tibble: 34 × 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 9.46e- 1 8.75e- 2 10.8 5.97e-27  
## 2 CLIENTNUM 1.50e-10 1.01e-10 1.49 1.37e- 1  
## 3 Customer\_Age -9.89e- 4 7.53e- 4 -1.31 1.89e- 1  
## 4 GenderM 7.36e- 2 1.34e- 2 5.48 4.48e- 8  
## 5 Dependent\_count -1.40e- 2 2.92e- 3 -4.79 1.69e- 6  
## 6 Education\_LevelDoctorate -5.17e- 2 2.10e- 2 -2.46 1.38e- 2  
## 7 Education\_LevelGraduate -1.67e- 3 1.36e- 2 -0.123 9.02e- 1  
## 8 Education\_LevelHigh School -1.05e- 3 1.44e- 2 -0.0730 9.42e- 1  
## 9 Education\_LevelPost-Graduate -1.64e- 2 2.05e- 2 -0.801 4.23e- 1  
## 10 Education\_LevelUneducated -8.74e- 3 1.52e- 2 -0.575 5.66e- 1  
## # … with 24 more rows

# Performance metrics on training data  
glance(lm\_fit)

## # A tibble: 1 × 12  
## r.squ…¹ adj.r…² sigma stati…³ p.value df logLik AIC BIC devia…⁴ df.re…⁵  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>  
## 1 0.359 0.356 0.271 94.1 0 32 -592. 1253. 1477. 394. 5364  
## # … with 1 more variable: nobs <int>, and abbreviated variable names  
## # ¹​r.squared, ²​adj.r.squared, ³​statistic, ⁴​deviance, ⁵​df.residual

We can also use the vip() function to plot the variable importance for each predictor in our model. The importance value is determined based on the F-statistics and estimate coefficents in our trained model object.

vip(lm\_fit)



predict(lm\_fit, new\_data = Bank\_churners)

## Warning in predict.lm(object = object$fit, newdata = new\_data, type =  
## "response"): prediction from a rank-deficient fit may be misleading

## # A tibble: 5,998 × 1  
## .pred  
## <dbl>  
## 1 2.16  
## 2 2.62  
## 3 2.17  
## 4 2.07  
## 5 2.33  
## 6 1.81  
## 7 2.04  
## 8 1.82  
## 9 2.18  
## 10 1.93  
## # … with 5,988 more rows

ggplot(data = Bank\_churners,  
 mapping = aes(x = Total\_Trans\_Amt, y = Total\_Trans\_Ct)) +  
 geom\_point(color = '#006EA1') +  
 geom\_abline(intercept = 0, slope = 1, color = 'orange') +  
 labs(title = 'Linear Regression Results - Total\_Trans\_Ct & Total\_Trans\_Amt',  
 x = 'Total\_Trans\_Amt',  
 y = 'Total\_Trans\_Ct')

