A3_DA3_JH_hope

Jana Hochel

2023-02-26

FEATURE ENGINEERING

Most companies did not grow. Median growth was about 0%. Depending on feature engineering, I have selected fast growing companies as those growing more than 7.89% (75th percentile).

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   -1.00     0.00     0.00     3.92     0.05 80795.00
```

OLS

I have ran OLS to get an idea what the coefficients might be. It seems that most variables coefficient is insignificant. Only sales and profit seems important.

```
Call:
lm(formula = formula(paste0("fastgrowing ~", paste0(X1, collapse = " + "))),
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9386 -0.3529 -0.2668  0.5633  1.0572

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.432692   0.010321  41.925  < 2e-16 ***
sales_mil_log         0.016733   0.002271   7.368 1.77e-13 ***
sales_mil_log_sq     -0.001228   0.000305  -4.027 5.66e-05 ***
d1_sales_mil_log_mod -0.355650   0.005648 -62.970  < 2e-16 ***
profit_loss_year_pl   0.053959   0.006647   8.118 4.87e-16 ***
ind2_cat27           -0.018084   0.015870  -1.139   0.2545
ind2_cat28           -0.018153   0.012261  -1.481   0.1387
ind2_cat29           -0.034586   0.020597  -1.679   0.0931 .
ind2_cat30           -0.035867   0.026152  -1.372   0.1702
ind2_cat33           -0.015332   0.012258  -1.251   0.2110
ind2_cat55            0.000607   0.011891   0.051   0.9593
ind2_cat56           -0.024793   0.010347  -2.396   0.0166 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4509 on 40170 degrees of freedom
```

```
Multiple R-squared:  0.09293,   Adjusted R-squared:  0.09268
F-statistic: 374.1 on 11 and 40170 DF,  p-value: < 2.2e-16


Call:
glm(formula = formula(paste0("fastgrowing ~", paste0(X1, collapse = " + "))),
    family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3813  -0.9195  -0.7381   1.2579   2.6165

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.252959   0.049524  -5.108 3.26e-07 ***
sales_mil_log         0.070572   0.011035   6.395 1.60e-10 ***
sales_mil_log_sq     -0.011522   0.001643  -7.012 2.36e-12 ***
d1_sales_mil_log_mod -2.073612   0.039330 -52.724  < 2e-16 ***
profit_loss_year_pl   0.305202   0.035203   8.670  < 2e-16 ***
ind2_cat27           -0.080070   0.076848  -1.042   0.2974
ind2_cat28           -0.091920   0.059294  -1.550   0.1211
ind2_cat29           -0.147109   0.099070  -1.485   0.1376
ind2_cat30           -0.202124   0.132712  -1.523   0.1278
ind2_cat33           -0.070088   0.059484  -1.178   0.2387
ind2_cat55            0.010541   0.057878   0.182   0.8555
ind2_cat56           -0.098624   0.050192  -1.965   0.0494 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 51458  on 40181  degrees of freedom
Residual deviance: 47228  on 40170  degrees of freedom
AIC: 47252

Number of Fisher Scoring iterations: 4


Call:
glm(formula = formula(paste0("fastgrowing ~", paste0(X2, collapse = " + "))),
    family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5384  -0.8670  -0.6775   1.1529   2.7549

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.991802   0.065894 -15.051  < 2e-16 ***
sales_mil_log         0.037006   0.011959   3.094  0.00197 **
sales_mil_log_sq     -0.014994   0.001733  -8.654  < 2e-16 ***
d1_sales_mil_log_mod -2.068675   0.039414 -52.486  < 2e-16 ***
profit_loss_year_pl   0.224226   0.038427   5.835 5.38e-09 ***
fixed_assets_bs       0.145887   0.037346   3.906 9.37e-05 ***
```

Table 1: Average Marginal Effects (dy/dx) for Logit Model

Variable	Coefficient	dx/dy
age	0.053	0.010
curr_liab_bs	-0.205	-0.040
curr_liab_bs_flag_error	-0.449	-0.087
curr_liab_bs_flag_high	0.134	0.026
d1_sales_mil_log_mod	-2.069	-0.403
fixed_assets_bs	0.146	0.028
foreign_management	-0.056	-0.011
ind2_cat27	-0.071	-0.013
ind2_cat28	-0.069	-0.013
ind2_cat29	-0.042	-0.008
ind2_cat30	-0.121	-0.022
ind2_cat33	0.121	0.023
ind2_cat55	0.102	0.020
ind2_cat56	0.145	0.028
profit_loss_year_pl	0.224	0.044
sales_mil_log	0.037	0.007
sales_mil_log_sq	-0.015	-0.003
share_eq_bs	-0.027	-0.005

```
share_eq_bs             -0.026604   0.028550  -0.932  0.35143
curr_liab_bs            -0.205087   0.052148  -3.933 8.40e-05 ***
curr_liab_bs_flag_high   0.133941   0.046016   2.911  0.00361 **
curr_liab_bs_flag_error -0.448670   0.574237  -0.781  0.43461
age                      0.053465   0.001693  31.576  < 2e-16 ***
foreign_management      -0.056256   0.038169  -1.474  0.14051
ind2_cat27              -0.071059   0.078009  -0.911  0.36235
ind2_cat28              -0.068587   0.060190  -1.140  0.25449
ind2_cat29              -0.041673   0.101248  -0.412  0.68064
ind2_cat30              -0.121063   0.134753  -0.898  0.36897
ind2_cat33               0.120539   0.060731   1.985  0.04717 *
ind2_cat55               0.101888   0.060746   1.677  0.09349 .
ind2_cat56               0.144606   0.052798   2.739  0.00617 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 51458  on 40181  degrees of freedom
Residual deviance: 45968  on 40163  degrees of freedom
AIC: 46006

Number of Fisher Scoring iterations: 4



Call:
lm(formula = formula(paste0("fastgrowing ~", paste0(X4, collapse = " + "))),
    data = data)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-1.09546 -0.03297 -0.00873  0.03959  1.04463

Coefficients: (2 not defined because of singularities)
                                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)                     8.910e-01  3.365e-02   26.479  < 2e-16 ***
sales_mil_log                  -1.084e-03  1.542e-03   -0.703 0.482131
sales_mil_log_sq                3.149e-04  1.857e-04    1.696 0.089934 .
age                             8.923e-04  5.873e-04    1.519 0.128727
age2                           -2.267e-06  2.401e-05   -0.094 0.924791
new                            -8.810e-01  2.530e-03 -348.226  < 2e-16 ***
m_region_locEast               -6.551e-04  2.978e-03   -0.220 0.825888
m_region_locWest               -2.160e-03  3.423e-03   -0.631 0.528002
urban_m2                        8.346e-03  3.244e-03    2.573 0.010090 *
urban_m3                        5.629e-03  3.007e-03    1.872 0.061215 .
total_assets_bs                -5.929e-10  3.400e-10   -1.744 0.081213 .
fixed_assets_bs                 1.860e-02  2.779e-02    0.669 0.503287
liq_assets_bs                  -7.059e-03  4.935e-03   -1.430 0.152664
curr_assets_bs                  1.649e-02  2.752e-02    0.599 0.548901
share_eq_bs                     2.785e-03  4.067e-03    0.685 0.493479
subscribed_cap_bs              -4.766e-03  5.216e-03   -0.914 0.360822
intang_assets_bs                       NA         NA       NA       NA
extra_exp_pl                   -1.472e-01  1.081e-01   -1.362 0.173363
extra_inc_pl                    2.744e-01  9.717e-02    2.824 0.004742 **
extra_profit_loss_pl           -2.684e-01  9.378e-02   -2.862 0.004207 **
inc_bef_tax_pl                  1.568e-02  2.304e-02    0.681 0.496106
inventories_pl                 -1.626e-02  8.393e-03   -1.937 0.052695 .
material_exp_pl                -8.517e-03  7.157e-03   -1.190 0.234066
profit_loss_year_pl            -2.947e-02  2.257e-02   -1.306 0.191631
personnel_exp_pl                1.275e-02  7.259e-03    1.756 0.079017 .
extra_profit_loss_pl_quad       3.407e-02  6.740e-02    0.505 0.613227
inc_bef_tax_pl_quad            -5.097e-03  2.630e-02   -0.194 0.846322
profit_loss_year_pl_quad       -3.337e-02  2.817e-02   -1.185 0.236118
share_eq_bs_quad                2.929e-03  5.095e-03    0.575 0.565348
extra_profit_loss_pl_flag_low   2.036e-02  1.416e-01    0.144 0.885718
inc_bef_tax_pl_flag_low         4.691e-02  5.003e-02    0.938 0.348489
profit_loss_year_pl_flag_low   -1.631e-02  5.030e-02   -0.324 0.745810
share_eq_bs_flag_low            6.639e-03  7.400e-03    0.897 0.369661
extra_exp_pl_flag_high         -8.381e-02  1.231e-01   -0.681 0.496131
extra_inc_pl_flag_high          3.655e-02  1.075e-01    0.340 0.733852
inventories_pl_flag_high        7.458e-04  1.018e-02    0.073 0.941584
material_exp_pl_flag_high       8.018e-03  4.366e-03    1.836 0.066332 .
personnel_exp_pl_flag_high      1.380e-02  7.704e-03    1.792 0.073211 .
curr_liab_bs_flag_high          2.211e-03  3.990e-03    0.554 0.579450
liq_assets_bs_flag_high        -1.472e-02  1.319e-01   -0.112 0.911159
subscribed_cap_bs_flag_high     3.158e-03  5.079e-03    0.622 0.534130
extra_profit_loss_pl_flag_high -9.628e-02  1.143e-01   -0.842 0.399537
inc_bef_tax_pl_flag_high       -2.349e-03  3.025e-02   -0.078 0.938122
profit_loss_year_pl_flag_high   6.954e-02  3.441e-02    2.021 0.043261 *
share_eq_bs_flag_high          -9.192e-03  1.179e-02   -0.780 0.435579
extra_exp_pl_flag_error         3.685e-02  9.518e-02    0.387 0.698607
extra_inc_pl_flag_error        -6.686e-02  1.078e-01   -0.620 0.534992
inventories_pl_flag_error      -1.059e-01  1.323e-01   -0.801 0.423395
material_exp_pl_flag_error     -1.086e-01  9.559e-02   -1.136 0.255963
```

```
personnel_exp_pl_flag_error     7.648e-02  9.523e-02    0.803 0.421909
curr_liab_bs_flag_error        -1.470e-03  5.450e-02   -0.027 0.978489
liq_assets_bs_flag_error        2.644e-02  3.548e-02    0.745 0.456164
subscribed_cap_bs_flag_error           NA         NA       NA       NA
extra_profit_loss_pl_flag_zero -1.132e-02  3.235e-03   -3.498 0.000469 ***
inc_bef_tax_pl_flag_zero        7.594e-03  2.741e-02    0.277 0.781718
profit_loss_year_pl_flag_zero   2.699e-03  7.127e-03    0.379 0.704939
share_eq_bs_flag_zero          -1.347e-01  8.797e-02   -1.531 0.125801
d1_sales_mil_log_mod           -6.085e-01  3.887e-03 -156.564  < 2e-16 ***
d1_sales_mil_log_mod_sq        -4.010e-01  4.575e-03  -87.652  < 2e-16 ***
flag_low_d1_sales_mil_log      -3.590e-01  1.318e-02  -27.244  < 2e-16 ***
flag_high_d1_sales_mil_log      9.280e-01  1.420e-02   65.340  < 2e-16 ***
female                          1.051e-03  2.723e-03    0.386 0.699482
ceo_age                         1.526e-04  1.137e-04    1.343 0.179397
flag_high_ceo_age               2.695e-03  1.404e-02    0.192 0.847811
flag_low_ceo_age               -1.980e-02  9.862e-03   -2.008 0.044700 *
flag_miss_ceo_age              -1.523e-02  5.105e-03   -2.983 0.002860 **
ceo_count                       5.750e-03  2.156e-03    2.667 0.007661 **
labor_avg_mod                   1.713e-03  3.917e-04    4.373 1.23e-05 ***
flag_miss_labor_avg            -1.275e-02  3.401e-03   -3.749 0.000178 ***
foreign_management              1.157e-02  4.846e-03    2.388 0.016950 *
balsheet_flag                   2.703e-01  1.984e-02   13.624  < 2e-16 ***
balsheet_length                -9.525e-05  4.671e-05   -2.039 0.041421 *
balsheet_notfullyear            6.249e-02  8.897e-03    7.024 2.20e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2152 on 40111 degrees of freedom
Multiple R-squared:  0.7937,    Adjusted R-squared:  0.7933
F-statistic:  2204 on 70 and 40111 DF,  p-value: < 2.2e-16


Call:
glm(formula = formula(paste0("fastgrowing ~", paste0(X4, collapse = " + "))),
    family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.3697  -0.1611  -0.1164   0.0000   3.5703

Coefficients: (2 not defined because of singularities)
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  1.634e+01  1.059e+00  15.422  < 2e-16 ***
sales_mil_log               -2.560e-01  4.929e-02  -5.193 2.07e-07 ***
sales_mil_log_sq            -7.466e-03  4.488e-03  -1.664  0.09618 .
age                         -2.677e-02  1.548e-02  -1.730  0.08363 .
age2                         1.398e-03  6.124e-04   2.282  0.02246 *
new                         -2.072e+01  6.296e-01 -32.910  < 2e-16 ***
m_region_locEast             3.018e-02  9.328e-02   0.324  0.74627
m_region_locWest            -7.545e-02  1.090e-01  -0.692  0.48870
urban_m2                     2.358e-01  1.021e-01   2.309  0.02093 *
urban_m3                     2.594e-01  9.498e-02   2.731  0.00632 **
total_assets_bs             -7.704e-10  5.190e-09  -0.148  0.88199
fixed_assets_bs             -7.294e-01  7.721e-01  -0.945  0.34484
```

```
liq_assets_bs                -3.033e-01  1.399e-01  -2.168  0.03015 *
curr_assets_bs               -4.941e-01  7.617e-01  -0.649  0.51655
share_eq_bs                   1.914e-01  1.278e-01   1.497  0.13429
subscribed_cap_bs            -2.866e-01  1.589e-01  -1.804  0.07124 .
intang_assets_bs                     NA         NA      NA       NA
extra_exp_pl                 -4.467e-01  2.165e+00  -0.206  0.83649
extra_inc_pl                  4.610e+00  2.111e+00   2.184  0.02899 *
extra_profit_loss_pl         -3.500e+00  1.983e+00  -1.765  0.07757 .
inc_bef_tax_pl                5.348e-01  6.773e-01   0.790  0.42976
inventories_pl               -2.003e-01  2.635e-01  -0.760  0.44711
material_exp_pl              -1.241e-01  2.024e-01  -0.613  0.53984
profit_loss_year_pl          -8.831e-01  6.678e-01  -1.322  0.18606
personnel_exp_pl             -2.406e-01  1.951e-01  -1.233  0.21757
extra_profit_loss_pl_quad    -1.219e+00  1.425e+00  -0.855  0.39243
inc_bef_tax_pl_quad          -2.856e-01  7.473e-01  -0.382  0.70229
profit_loss_year_pl_quad      2.206e-01  7.949e-01   0.277  0.78143
share_eq_bs_quad              1.486e-01  1.614e-01   0.921  0.35722
extra_profit_loss_pl_flag_low 2.119e+00  3.419e+00   0.620  0.53534
inc_bef_tax_pl_flag_low       1.144e+00  1.382e+00   0.828  0.40774
profit_loss_year_pl_flag_low -1.091e+00  1.386e+00  -0.787  0.43129
share_eq_bs_flag_low          4.645e-01  2.284e-01   2.034  0.04199 *
extra_exp_pl_flag_high       -3.454e+00  3.060e+00  -1.129  0.25898
extra_inc_pl_flag_high       -7.977e-02  3.030e+00  -0.026  0.97900
inventories_pl_flag_high      2.920e-02  2.908e-01   0.100  0.92001
material_exp_pl_flag_high     8.159e-02  1.248e-01   0.654  0.51327
personnel_exp_pl_flag_high    4.600e-01  1.778e-01   2.587  0.00968 **
curr_liab_bs_flag_high        2.192e-01  1.214e-01   1.805  0.07108 .
liq_assets_bs_flag_high      -1.097e+01  1.272e+03  -0.009  0.99312
subscribed_cap_bs_flag_high   1.574e-01  1.409e-01   1.117  0.26408
extra_profit_loss_pl_flag_high 2.187e-01 3.138e+00   0.070  0.94443
inc_bef_tax_pl_flag_high     -5.757e-01  8.879e-01  -0.648  0.51675
profit_loss_year_pl_flag_high 1.423e+00  9.641e-01   1.476  0.14006
share_eq_bs_flag_high        -1.250e-01  3.487e-01  -0.359  0.71992
extra_exp_pl_flag_error      -2.451e+00  1.897e+01  -0.129  0.89721
extra_inc_pl_flag_error      -6.505e+00  2.071e+01  -0.314  0.75342
inventories_pl_flag_error    -1.360e+01  1.261e+03  -0.011  0.99139
material_exp_pl_flag_error   -2.639e+00  2.011e+00  -1.312  0.18950
personnel_exp_pl_flag_error   1.073e+00  1.976e+00   0.543  0.58691
curr_liab_bs_flag_error      -3.864e+00  1.107e+01  -0.349  0.72694
liq_assets_bs_flag_error      7.021e-01  7.423e-01   0.946  0.34422
subscribed_cap_bs_flag_error         NA         NA      NA       NA
extra_profit_loss_pl_flag_zero -2.765e-01 1.100e-01  -2.513  0.01196 *
inc_bef_tax_pl_flag_zero     -2.729e-01  7.112e-01  -0.384  0.70118
profit_loss_year_pl_flag_zero 3.347e-01  2.366e-01   1.414  0.15723
share_eq_bs_flag_zero        -1.415e+01  9.686e+02  -0.015  0.98834
d1_sales_mil_log_mod         -4.097e+01  1.597e+00 -25.651  < 2e-16 ***
d1_sales_mil_log_mod_sq      -3.498e+01  1.381e+00 -25.334  < 2e-16 ***
flag_low_d1_sales_mil_log     8.194e-01  4.572e-01   1.792  0.07310 .
flag_high_d1_sales_mil_log    1.074e+02  1.024e+02   1.049  0.29425
female                       -2.500e-01  8.745e-02  -2.859  0.00425 **
ceo_age                       1.629e-03  3.595e-03   0.453  0.65049
flag_high_ceo_age             6.396e-02  4.448e-01   0.144  0.88565
flag_low_ceo_age             -5.441e-01  2.867e-01  -1.898  0.05776 .
flag_miss_ceo_age            -1.455e-01  1.594e-01  -0.913  0.36124
```

```
ceo_count                      1.977e-01  7.028e-02   2.813  0.00491 **
labor_avg_mod                  2.603e-02  8.268e-03   3.149  0.00164 **
flag_miss_labor_avg           -5.614e-01  1.007e-01  -5.575 2.48e-08 ***
foreign_management            -1.087e-02  1.554e-01  -0.070  0.94422
balsheet_flag                  2.751e+00  2.844e-01   9.673  < 2e-16 ***
balsheet_length               -7.063e-04  7.756e-04  -0.911  0.36250
balsheet_notfullyear           1.662e+00  1.685e-01   9.867  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 51458  on 40181  degrees of freedom
Residual deviance:  6623  on 40111  degrees of freedom
AIC: 6765

Number of Fisher Scoring iterations: 15
```

Average Marginal Effect

```
[1] 32146   123
```

```
[1] 8036  123
```

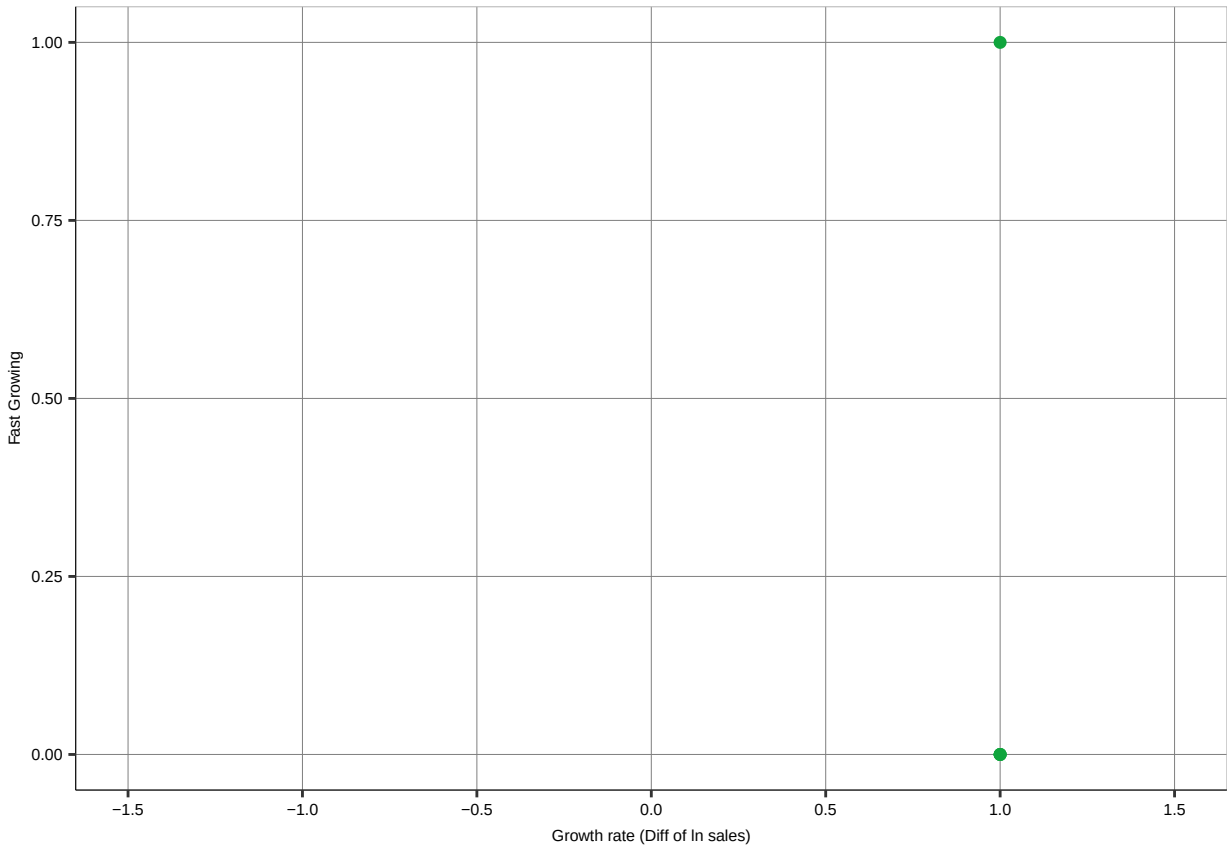
```
data$fastgrowing
       n  missing distinct     Info      Sum     Mean      Gmd
   40182        0        2    0.672    13618   0.3389   0.4481
```

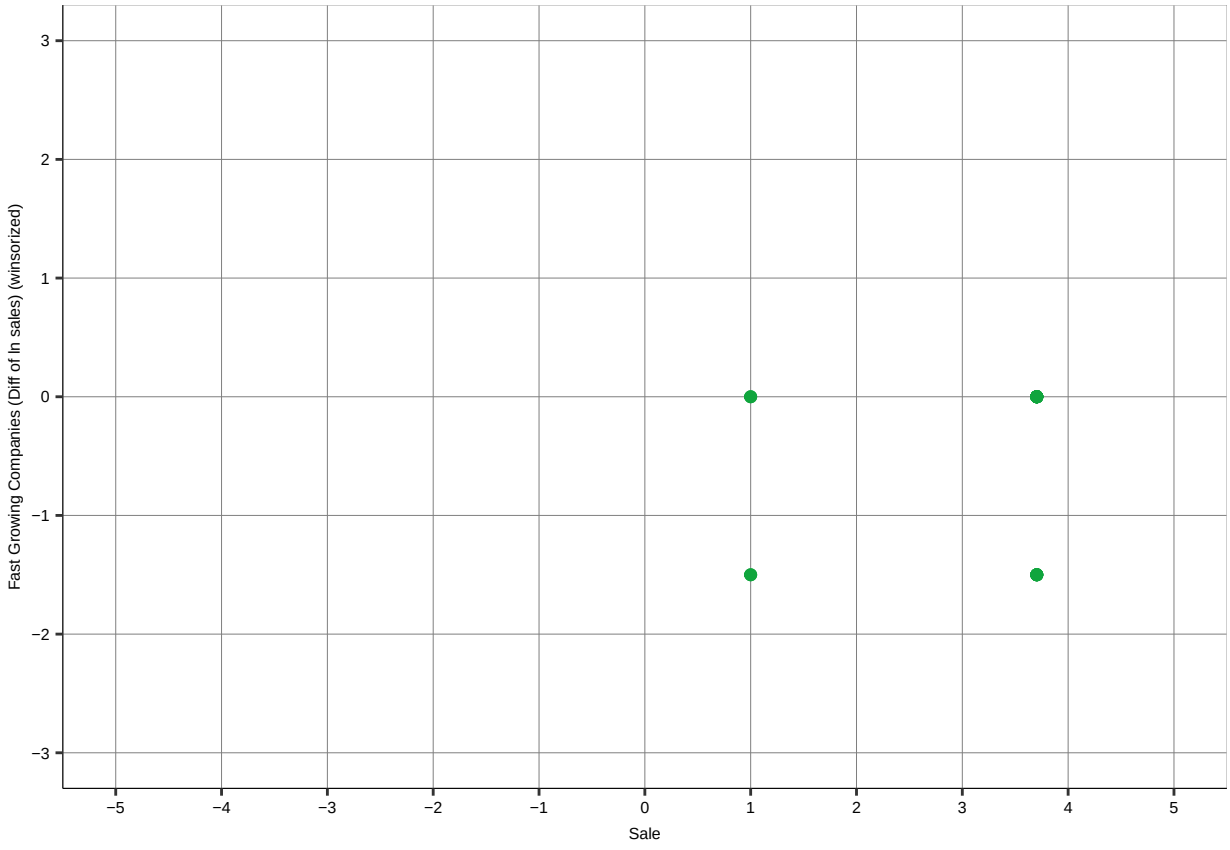
```
data_train$fastgrowing
       n  missing distinct     Info      Sum     Mean      Gmd
   32146        0        2    0.672    10898    0.339   0.4482
```

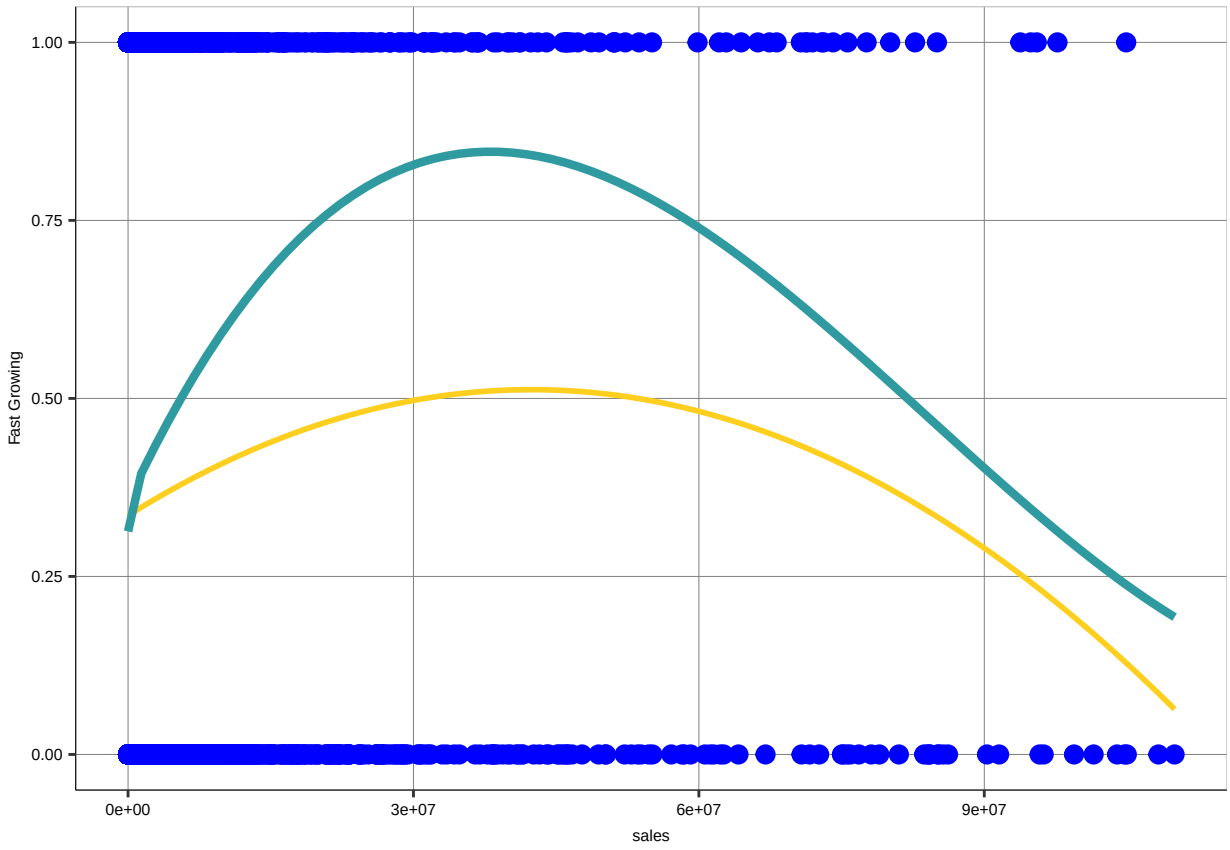
```
data_holdout$fastgrowing
       n  missing distinct     Info      Sum     Mean      Gmd
    8036        0        2    0.672     2720   0.3385   0.4479
```

Table 2: Average Marginal Effects (dy/dx) for Logit Model

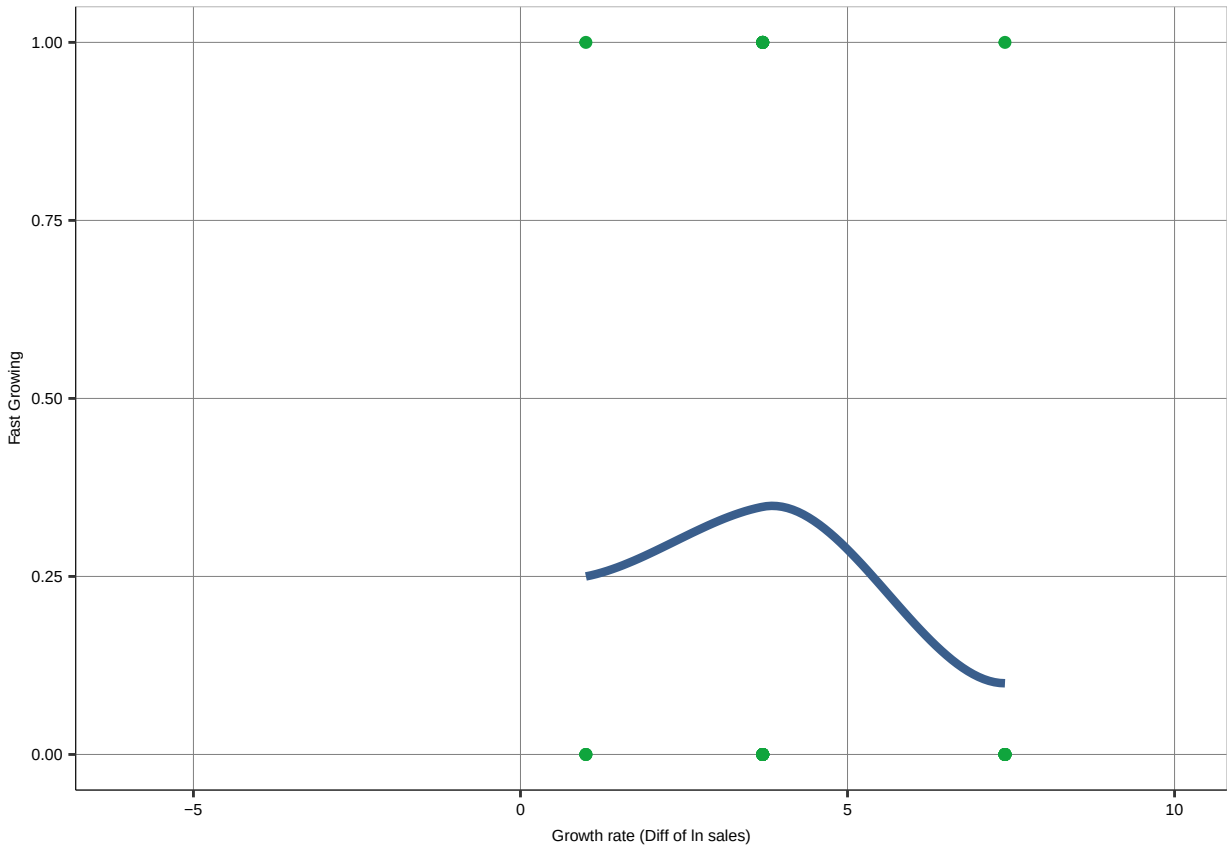
Variable	Coefficient	SE	dx/dy
age	-0.027	0.015	-0.001
age2	0.001	0.001	0.000
balsheet_flag	2.751	0.284	0.059
balsheet_length	-0.001	0.001	0.000
balsheet_notfullyear	1.662	0.168	0.036
ceo_age	0.002	0.004	0.000
ceo_count	0.198	0.070	0.004
curr_assets_bs	-0.494	0.762	-0.011
curr_liab_bs_flag_error	-3.864	11.066	-0.084
curr_liab_bs_flag_high	0.219	0.121	0.005
d1_sales_mil_log_mod	-40.965	1.597	-0.885
d1_sales_mil_log_mod_sq	-34.981	1.381	-0.756
extra_exp_pl	-0.447	2.165	-0.010
extra_exp_pl_flag_error	-2.451	18.969	-0.053
extra_exp_pl_flag_high	-3.454	3.060	-0.075
extra_inc_pl	4.610	2.111	0.100
extra_inc_pl_flag_error	-6.505	20.707	-0.141
extra_inc_pl_flag_high	-0.080	3.030	-0.002
extra_profit_loss_pl	-3.500	1.983	-0.076
extra_profit_loss_pl_flag_high	0.219	3.138	0.005
extra_profit_loss_pl_flag_low	2.119	3.419	0.046
extra_profit_loss_pl_flag_zero	-0.277	0.110	-0.006
extra_profit_loss_pl_quad	-1.219	1.425	-0.026
female	-0.250	0.087	-0.005
fixed_assets_bs	-0.729	0.772	-0.016
flag_high_ceo_age	0.064	0.445	0.001
flag_high_d1_sales_mil_log	107.430	102.428	2.322
flag_low_ceo_age	-0.544	0.287	-0.012
flag_low_d1_sales_mil_log	0.819	0.457	0.018
flag_miss_ceo_age	-0.145	0.159	-0.003
flag_miss_labor_avg	-0.561	0.101	-0.012
foreign_management	-0.011	0.155	0.000
inc_bef_tax_pl	0.535	0.677	0.012
inc_bef_tax_pl_flag_high	-0.576	0.888	-0.012
inc_bef_tax_pl_flag_low	1.144	1.382	0.025
inc_bef_tax_pl_flag_zero	-0.273	0.711	-0.006
inc_bef_tax_pl_quad	-0.286	0.747	-0.006
inventories_pl	-0.200	0.264	-0.004
inventories_pl_flag_error	-13.604	1261.145	-0.294
inventories_pl_flag_high	0.029	0.291	0.001
labor_avg_mod	0.026	0.008	0.001
liq_assets_bs	-0.303	0.140	-0.007
liq_assets_bs_flag_error	0.702	0.742	0.015
liq_assets_bs_flag_high	-10.970	1271.814	-0.237
m_region_locEast	0.030	0.093	0.001
m_region_locWest	-0.075	0.109	-0.002
material_exp_pl	-0.124	0.202	-0.003
material_exp_pl_flag_error	-2.639	2.011	-0.057
material_exp_pl_flag_high	0.082	0.125	0.002
new	-20.720	0.630	-0.448
personnel_exp_pl	-0.241	0.195	-0.005
personnel_exp_pl_flag_error	1.073	1.976	0.023
personnel_exp_pl_flag_high	0.460	0.178	0.010
profit_loss_year_pl	-0.883	0.668	-0.019
profit_loss_year_pl_flag_high	1.423	0.964	0.031

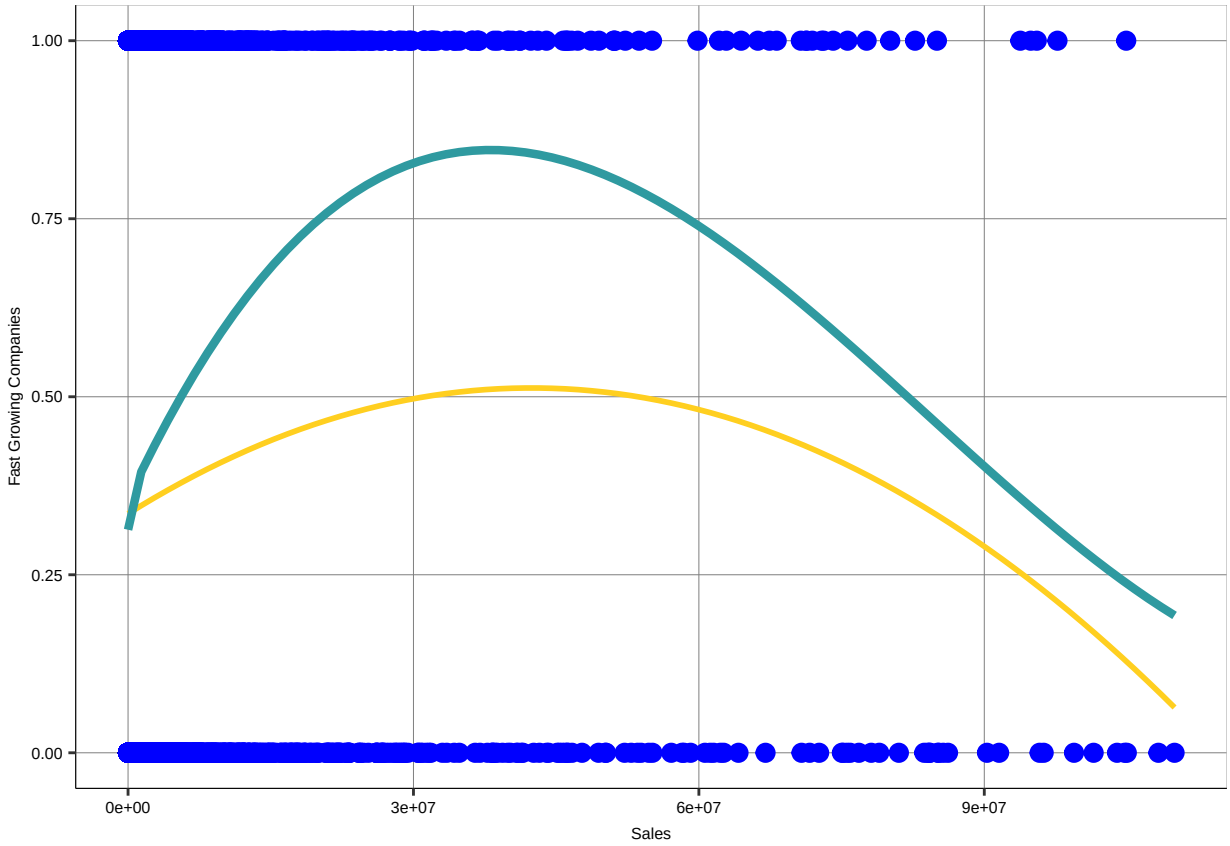

1.00
0.75
0.50
0.25
0.00
Fast Growing
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Growth rate (Diff of ln sales)

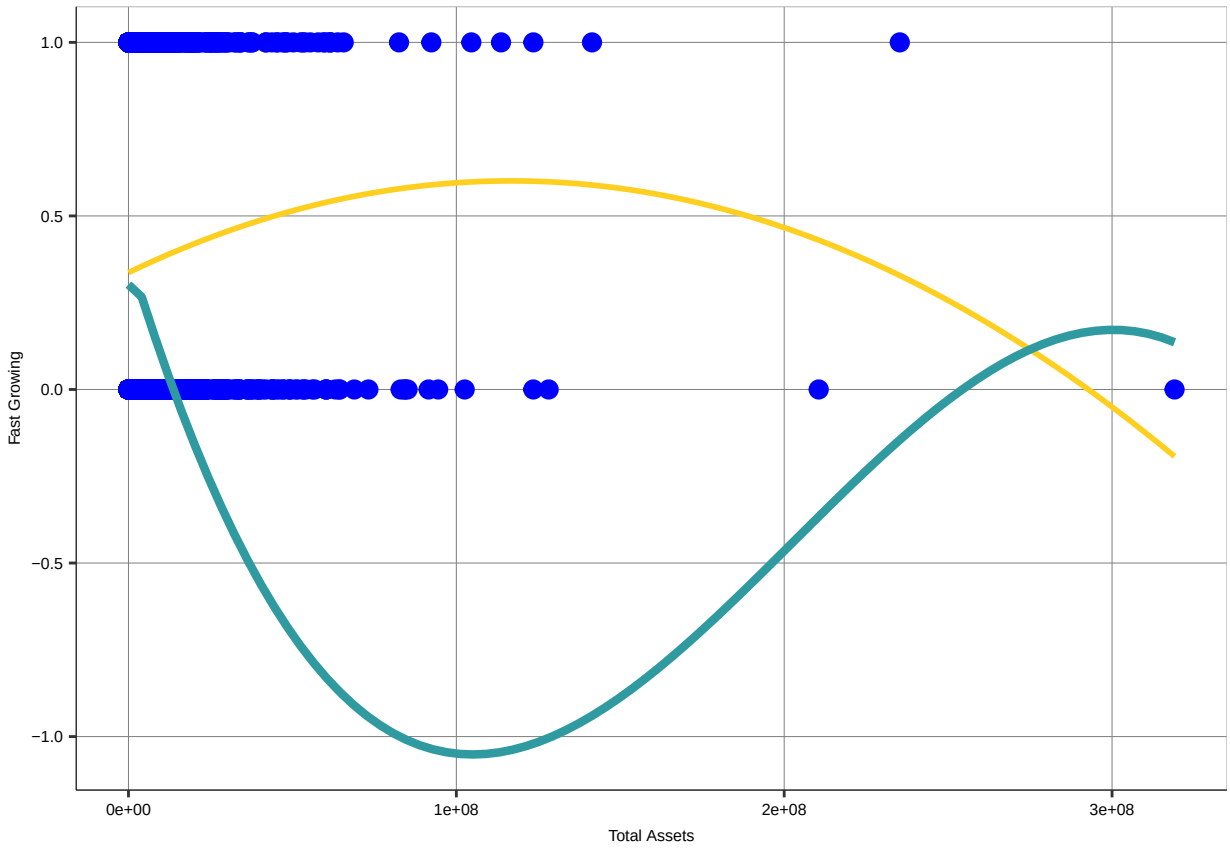

3
2
1
0
−1
−2
−3
Fast Growing Companies (Diff of ln sales) (winsorized)
−5
−4
−3
−2
−1
0
1
2
3
4
5
Sale




```
Call:
lm(formula = fastgrowing ~ sales_mil_log + sales_mil_log_sq,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4606 -0.3424 -0.3229  0.6477  0.7272

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.3777130  0.0045839  82.400  < 2e-16 ***
sales_mil_log    0.0150917  0.0022450   6.722 1.81e-11 ***
sales_mil_log_sq 0.0005425  0.0003063   1.771   0.0765 .
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.4727 on 40179 degrees of freedom
Multiple R-squared:  0.00263,   Adjusted R-squared:  0.00258
F-statistic: 52.97 on 2 and 40179 DF,  p-value: < 2.2e-16
```


1.00
0.75
0.50
0.25
0.00
Fast Growing
−5
0
5
10
Growth rate (Diff of ln sales)


1.00
0.75
0.50
0.25
0.00
Fast Growing Companies
0e+00
3e+07
6e+07
9e+07
Sales




```
Call:
lm(formula = fastgrowing ~ sales + total_assets_bs + urban_m +
    COGS + female + ceo_count + profit_loss_year, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5388 -0.3815 -0.3618  0.6137  0.6861

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.527e-01  2.594e-02  13.594   <2e-16 ***
sales             2.972e-10  9.381e-10   0.317    0.751
total_assets_bs   5.106e-10  8.409e-10   0.607    0.544
urban_m2          1.459e-02  2.615e-02   0.558    0.577
urban_m3          1.983e-02  2.497e-02   0.794    0.427
COGS              1.310e-09  3.465e-09   0.378    0.705
female           -4.734e-02  3.598e-02  -1.316    0.188
ceo_count         8.583e-03  1.150e-02   0.746    0.455
profit_loss_year -1.068e-09  5.438e-09  -0.196    0.844
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.4855 on 2213 degrees of freedom
  (37960 observations deleted due to missingness)
Multiple R-squared:  0.002632,  Adjusted R-squared:  -0.0009734
```

```
F-statistic:  0.73 on 8 and 2213 DF,  p-value: 0.6651
```

PART I PREDICT PROBABILITIES

Predict probabilities Look at cross-validated performance and pick your favorite model

Logit model

```
   user  system elapsed
 235.89    1.75  239.73
```

```
       s1
 Min.   :-1.30189
 1st Qu.: 0.00000
 Median : 0.00000
 Mean   :-0.01552
 3rd Qu.: 0.00000
 Max.   : 0.00000
```

```
lasso_coeffs
```

```
 1  Variables      154  Observations
---------------------------------------------------------------------------
s1
       n  missing distinct     Info     Mean      Gmd
     154        0        4    0.057 -0.01552  0.03082

Value      -1.3018917 -0.9302874 -0.1573768  0.0000000
Frequency           1          1          1        151
Proportion      0.006      0.006      0.006      0.981
---------------------------------------------------------------------------
```

Random Forest model

```
[1] "character"
```

```
[1] 32146   123
```

```
[1] 8036  123
```

```
   user  system elapsed
 765.70    2.03  117.06
```

```
Random Forest

32146 samples
   20 predictor
```

```
No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 25715, 25718, 25717, 25717, 25717
Resampling results:

  RMSE      Rsquared      MAE
  312.1932  0.0002899016  8.936287

Tuning parameter 'mtry' was held constant at a value of 8
Tuning
 parameter 'splitrule' was held constant at a value of variance

Tuning parameter 'min.node.size' was held constant at a value of 50
```

	Length	Class	Mode
predictions	32146	-none-	numeric
num.trees	1	-none-	numeric
num.independent.variables	1	-none-	numeric
mtry	1	-none-	numeric
min.node.size	1	-none-	numeric
variable.importance	21	-none-	numeric
prediction.error	1	-none-	numeric
forest	7	ranger.forest	list
splitrule	1	-none-	character
treetype	1	-none-	character
r.squared	1	-none-	numeric
call	9	-none-	call
importance.mode	1	-none-	character
num.samples	1	-none-	numeric
replace	1	-none-	logical
xNames	21	-none-	character
problemType	1	-none-	character
tuneValue	3	data.frame	list
obsLevels	1	-none-	logical
param	1	-none-	list

PART III Classification forest

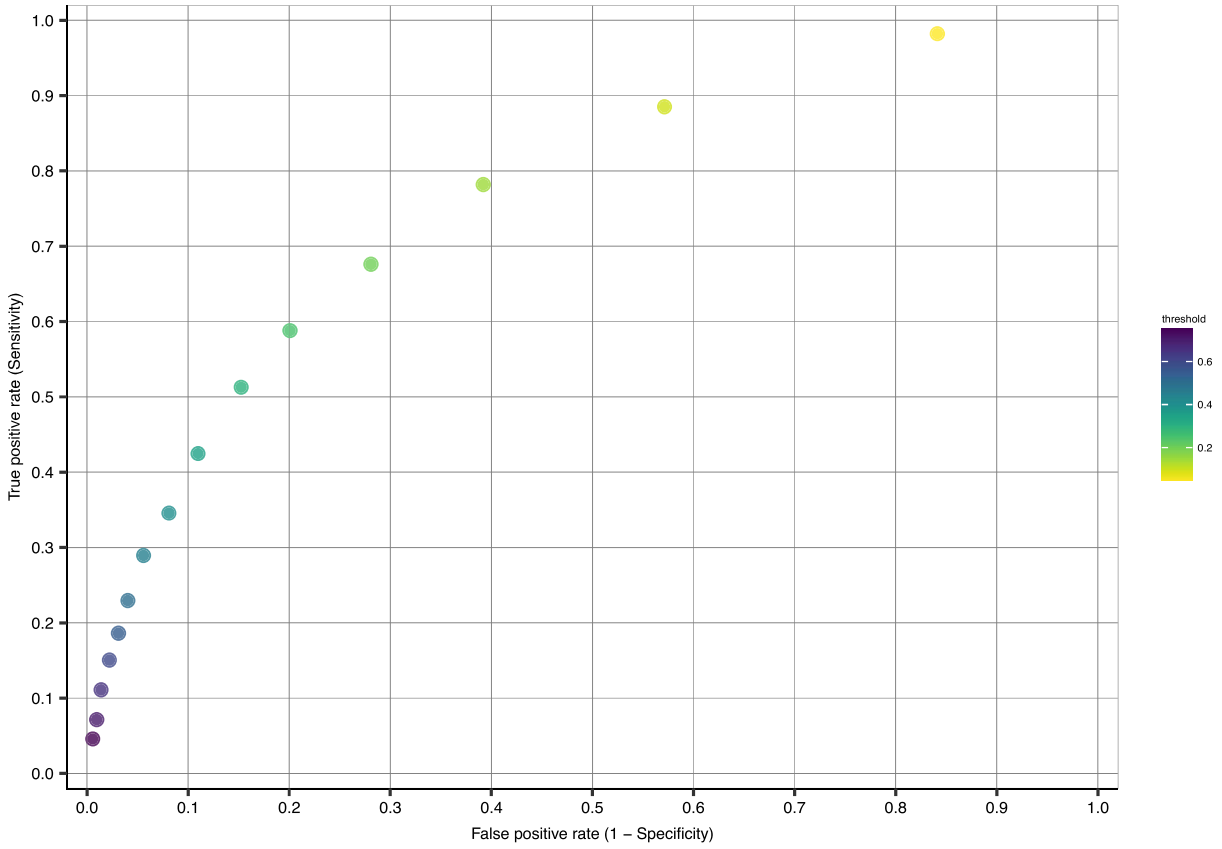
Show a confusion table (on a selected fold or holdout set) Discuss results, evaluate how useful your model may be

PART I No loss function

AUC is the Area Under the Curve of ROC function.
It is used to determine the probabilities of an event.
Our goal is to maximize AUC and minimize RMSE.

	Number.of.predictors	CV.RMSE	CV.AUC
X1	11	0.3760114	0.7332908
X2	18	0.3685874	0.7668024
X3	35	0.3682478	0.7671841
X4	79	0.3649985	0.7755887
X5	153	0.3658922	0.7716770
LASSO	89	0.3638972	0.7612783

```
[1] 0.3698712
```




```
Area under the curve: 0.7709
```

```
      Lagger      Fast Growing
       3505         302
```

```
Prediction          Lagger   Fast growing
      Lagger        2901      604
      Fast Growing   122      180
```

```
Prediction          Lagger     Fast Growing
     Lagger         2901      604
     Fast Growing    122      180
```

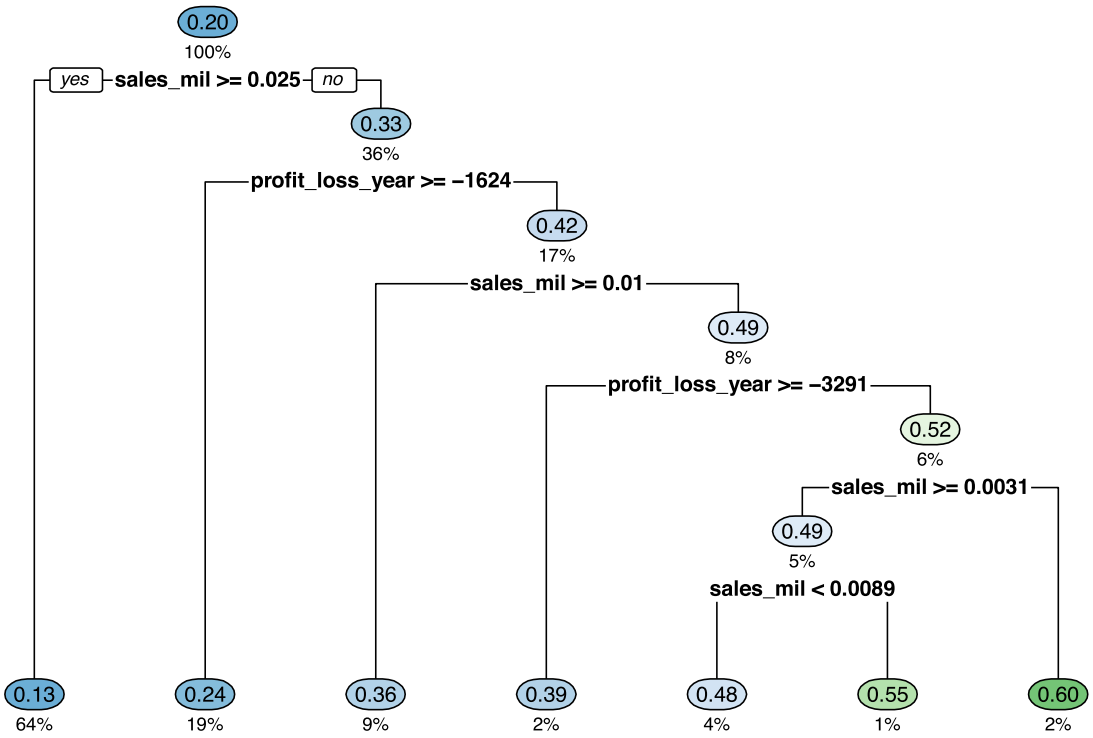
```
[1] 0.2014269
```

```
            Reference
Prediction
                   2183      255
                    840      529
```

```
[1] "best_logit_no_loss_pred" "default"
```

	Avg.of.optimal.thresholds	Threshold.for.Fold5	Avg.expected.loss	Expected.loss.for.Fold5
X1	0.0878080	0.0836493	0.7184334	0.7301379
X2	0.1068464	0.0954040	0.6519156	0.6539724
X3	0.0936328	0.0927424	0.6569064	0.6585686
X4	0.0971508	0.0815276	0.6457428	0.6454366
X5	0.0932000	0.0827964	0.6551330	0.6441234
LASSO	0.1048994	0.0866703	0.6813330	0.6772817

```
[1] 0.688994
```




```
+ Fold1: mtry=5, splitrule=gini, min.node.size=10
- Fold1: mtry=5, splitrule=gini, min.node.size=10
+ Fold1: mtry=6, splitrule=gini, min.node.size=10
- Fold1: mtry=6, splitrule=gini, min.node.size=10
+ Fold1: mtry=7, splitrule=gini, min.node.size=10
- Fold1: mtry=7, splitrule=gini, min.node.size=10
+ Fold1: mtry=5, splitrule=gini, min.node.size=15
- Fold1: mtry=5, splitrule=gini, min.node.size=15
+ Fold1: mtry=6, splitrule=gini, min.node.size=15
- Fold1: mtry=6, splitrule=gini, min.node.size=15
+ Fold1: mtry=7, splitrule=gini, min.node.size=15
- Fold1: mtry=7, splitrule=gini, min.node.size=15
+ Fold2: mtry=5, splitrule=gini, min.node.size=10
- Fold2: mtry=5, splitrule=gini, min.node.size=10
+ Fold2: mtry=6, splitrule=gini, min.node.size=10
- Fold2: mtry=6, splitrule=gini, min.node.size=10
+ Fold2: mtry=7, splitrule=gini, min.node.size=10
- Fold2: mtry=7, splitrule=gini, min.node.size=10
+ Fold2: mtry=5, splitrule=gini, min.node.size=15
- Fold2: mtry=5, splitrule=gini, min.node.size=15
+ Fold2: mtry=6, splitrule=gini, min.node.size=15
- Fold2: mtry=6, splitrule=gini, min.node.size=15
+ Fold2: mtry=7, splitrule=gini, min.node.size=15
- Fold2: mtry=7, splitrule=gini, min.node.size=15
+ Fold3: mtry=5, splitrule=gini, min.node.size=10
- Fold3: mtry=5, splitrule=gini, min.node.size=10
+ Fold3: mtry=6, splitrule=gini, min.node.size=10
- Fold3: mtry=6, splitrule=gini, min.node.size=10
+ Fold3: mtry=7, splitrule=gini, min.node.size=10
- Fold3: mtry=7, splitrule=gini, min.node.size=10
```

```
+ Fold3: mtry=5, splitrule=gini, min.node.size=15
- Fold3: mtry=5, splitrule=gini, min.node.size=15
+ Fold3: mtry=6, splitrule=gini, min.node.size=15
- Fold3: mtry=6, splitrule=gini, min.node.size=15
+ Fold3: mtry=7, splitrule=gini, min.node.size=15
- Fold3: mtry=7, splitrule=gini, min.node.size=15
+ Fold4: mtry=5, splitrule=gini, min.node.size=10
- Fold4: mtry=5, splitrule=gini, min.node.size=10
+ Fold4: mtry=6, splitrule=gini, min.node.size=10
- Fold4: mtry=6, splitrule=gini, min.node.size=10
+ Fold4: mtry=7, splitrule=gini, min.node.size=10
- Fold4: mtry=7, splitrule=gini, min.node.size=10
+ Fold4: mtry=5, splitrule=gini, min.node.size=15
- Fold4: mtry=5, splitrule=gini, min.node.size=15
+ Fold4: mtry=6, splitrule=gini, min.node.size=15
- Fold4: mtry=6, splitrule=gini, min.node.size=15
+ Fold4: mtry=7, splitrule=gini, min.node.size=15
- Fold4: mtry=7, splitrule=gini, min.node.size=15
+ Fold5: mtry=5, splitrule=gini, min.node.size=10
- Fold5: mtry=5, splitrule=gini, min.node.size=10
+ Fold5: mtry=6, splitrule=gini, min.node.size=10
- Fold5: mtry=6, splitrule=gini, min.node.size=10
+ Fold5: mtry=7, splitrule=gini, min.node.size=10
- Fold5: mtry=7, splitrule=gini, min.node.size=10
+ Fold5: mtry=5, splitrule=gini, min.node.size=15
- Fold5: mtry=5, splitrule=gini, min.node.size=15
+ Fold5: mtry=6, splitrule=gini, min.node.size=15
- Fold5: mtry=6, splitrule=gini, min.node.size=15
+ Fold5: mtry=7, splitrule=gini, min.node.size=15
- Fold5: mtry=7, splitrule=gini, min.node.size=15
Aggregating results
Selecting tuning parameters
Fitting mtry = 6, splitrule = gini, min.node.size = 15 on full training set

  mtry splitrule min.node.size  Accuracy     Kappa      RMSE  AccuracySD
1    5      gini            10 0.8213938 0.2637498 0.3560902 0.005469264
2    5      gini            15 0.8204086 0.2602668 0.3563651 0.004809608
3    6      gini            10 0.8213280 0.2702299 0.3562653 0.005022607
4    6      gini            15 0.8225756 0.2738490 0.3561898 0.005322257
5    7      gini            10 0.8225101 0.2785214 0.3563910 0.006475467
6    7      gini            15 0.8207371 0.2683172 0.3565271 0.005309215
     KappaSD      RMSESD
1 0.02503247 0.001786709
2 0.02404681 0.001953811
3 0.02307786 0.002060298
4 0.02227836 0.002191018
5 0.02859412 0.002328203
6 0.02658400 0.002076375
```

CV.RMSE	CV.AUC	Avg.of.optimal.thresholds	Threshold.for.Fold5	Avg.expected.loss	Expected.loss.for.Fold5
0.3561898	0.8044629	0.1165538	0.1106151	0.59052	0.5929087

```
[1] 0.3583408
```

```
[1] 0.8051229
```

```
[1] 0.6214867
```

PART II Classification forest

```
+ Fold1: mtry=5, splitrule=gini, min.node.size=10
- Fold1: mtry=5, splitrule=gini, min.node.size=10
+ Fold1: mtry=6, splitrule=gini, min.node.size=10
- Fold1: mtry=6, splitrule=gini, min.node.size=10
+ Fold1: mtry=7, splitrule=gini, min.node.size=10
- Fold1: mtry=7, splitrule=gini, min.node.size=10
+ Fold1: mtry=5, splitrule=gini, min.node.size=15
- Fold1: mtry=5, splitrule=gini, min.node.size=15
+ Fold1: mtry=6, splitrule=gini, min.node.size=15
- Fold1: mtry=6, splitrule=gini, min.node.size=15
+ Fold1: mtry=7, splitrule=gini, min.node.size=15
- Fold1: mtry=7, splitrule=gini, min.node.size=15
+ Fold2: mtry=5, splitrule=gini, min.node.size=10
- Fold2: mtry=5, splitrule=gini, min.node.size=10
+ Fold2: mtry=6, splitrule=gini, min.node.size=10
- Fold2: mtry=6, splitrule=gini, min.node.size=10
+ Fold2: mtry=7, splitrule=gini, min.node.size=10
- Fold2: mtry=7, splitrule=gini, min.node.size=10
+ Fold2: mtry=5, splitrule=gini, min.node.size=15
- Fold2: mtry=5, splitrule=gini, min.node.size=15
+ Fold2: mtry=6, splitrule=gini, min.node.size=15
- Fold2: mtry=6, splitrule=gini, min.node.size=15
+ Fold2: mtry=7, splitrule=gini, min.node.size=15
- Fold2: mtry=7, splitrule=gini, min.node.size=15
+ Fold3: mtry=5, splitrule=gini, min.node.size=10
- Fold3: mtry=5, splitrule=gini, min.node.size=10
+ Fold3: mtry=6, splitrule=gini, min.node.size=10
- Fold3: mtry=6, splitrule=gini, min.node.size=10
+ Fold3: mtry=7, splitrule=gini, min.node.size=10
- Fold3: mtry=7, splitrule=gini, min.node.size=10
+ Fold3: mtry=5, splitrule=gini, min.node.size=15
- Fold3: mtry=5, splitrule=gini, min.node.size=15
+ Fold3: mtry=6, splitrule=gini, min.node.size=15
- Fold3: mtry=6, splitrule=gini, min.node.size=15
+ Fold3: mtry=7, splitrule=gini, min.node.size=15
- Fold3: mtry=7, splitrule=gini, min.node.size=15
+ Fold4: mtry=5, splitrule=gini, min.node.size=10
- Fold4: mtry=5, splitrule=gini, min.node.size=10
+ Fold4: mtry=6, splitrule=gini, min.node.size=10
- Fold4: mtry=6, splitrule=gini, min.node.size=10
+ Fold4: mtry=7, splitrule=gini, min.node.size=10
- Fold4: mtry=7, splitrule=gini, min.node.size=10
+ Fold4: mtry=5, splitrule=gini, min.node.size=15
- Fold4: mtry=5, splitrule=gini, min.node.size=15
+ Fold4: mtry=6, splitrule=gini, min.node.size=15
- Fold4: mtry=6, splitrule=gini, min.node.size=15
```

```
+ Fold4: mtry=7, splitrule=gini, min.node.size=15
- Fold4: mtry=7, splitrule=gini, min.node.size=15
+ Fold5: mtry=5, splitrule=gini, min.node.size=10
- Fold5: mtry=5, splitrule=gini, min.node.size=10
+ Fold5: mtry=6, splitrule=gini, min.node.size=10
- Fold5: mtry=6, splitrule=gini, min.node.size=10
+ Fold5: mtry=7, splitrule=gini, min.node.size=10
- Fold5: mtry=7, splitrule=gini, min.node.size=10
+ Fold5: mtry=5, splitrule=gini, min.node.size=15
- Fold5: mtry=5, splitrule=gini, min.node.size=15
+ Fold5: mtry=6, splitrule=gini, min.node.size=15
- Fold5: mtry=6, splitrule=gini, min.node.size=15
+ Fold5: mtry=7, splitrule=gini, min.node.size=15
- Fold5: mtry=7, splitrule=gini, min.node.size=15
Aggregating results
Selecting tuning parameters
Fitting mtry = 6, splitrule = gini, min.node.size = 15 on full training set

[1] 1.525348

               Number.of.predictors   CV.RMSE    CV.AUC CV.threshold
Logit X1                         11 0.3760114 0.7332908    0.0878080
Logit X4                         79 0.3649985 0.7755887    0.0971508
Logit LASSO                      89 0.3638972 0.7612783    0.1048994
RF probability                   36 0.3561898 0.8044629    0.1165538
               CV.expected.Loss
Logit X1              0.7184334
Logit X4              0.6457428
Logit LASSO           0.6813330
RF probability        0.5905200
```

PART III

There were very few fast growing companies and many extreme values.
Logit model performed better than the others (0.70 probability).

Sales and Profit have the biggest impact on the ability of the firm to grow.