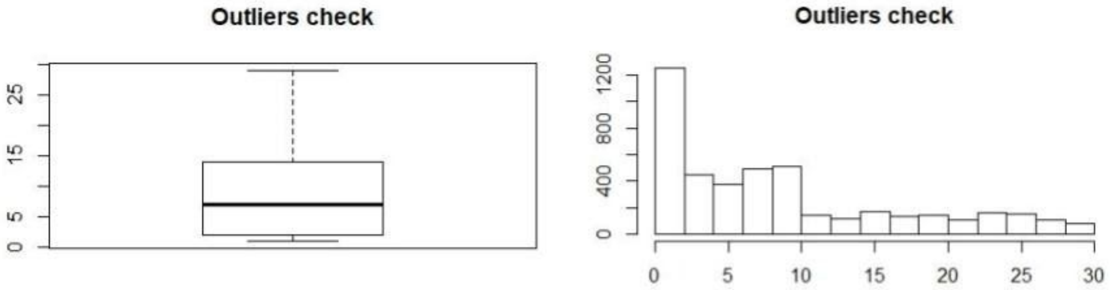
HR Analytics Case Study

1. Data import and preparation:

Firstly all data sets was imported: general data, employee survey, manager survey and both time stamps files.

Then, using boxplots and histograms I have checked the presence of any outliers. Example of that check is shown below on Pic. 1.



Pic.1: Outliers check for variable 'Distance from home'

No outliers was found in the data.

After that I had to adjust data sets with employees log in and log out times. It was in a form of panel data, covering almost 6 months period which resulted in table with 262 columns.
I decide to compute average log in time and average log out time. I was also able to create a feature which contained information about the amount of absent days for each employee.

Categorical variables were converted to dummy variables, numeric variables was scaled, rows with missing data (110 rows) was deleted from the dataset.

2. Estimation of Logistic Regression model

Dataset was split into train and test set, and on train set logistic regression model was estimated. After that I performed variables selection using stepwise method. Model summary is shown below:

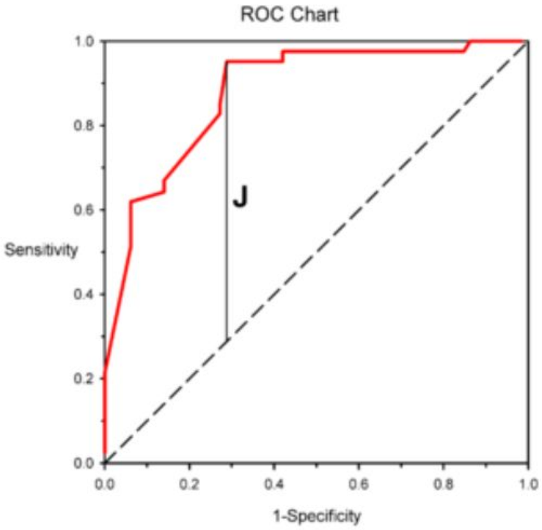
```
Call:
glm(formula = y ~ EnvironmentSatisfaction + JobSatisfaction +
    WorkLifeBalance + Age + NumCompaniesWorked + TotalWorkingYears +
    TrainingTimesLastYear + YearsSinceLastPromotion + YearsWithCurrManager +
    emp_outtime_int_avg + `BusinessTravel - Non-Travel` + `BusinessTravel - Travel_Frequently` +
    `Department - Human Resources` + `JobRole - Research Director` +
    `JobRole - Research Scientist` + `JobRole - Sales Executive` +
    `MaritalStatus - Divorced` + `MaritalStatus - Married`, family = binomial(link = "logit"),
    data = d_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8028  -0.5645  -0.3509  -0.1700   3.7679

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -1.93728    0.11952 -16.208  < 2e-16 ***
EnvironmentSatisfaction             -0.38223    0.05424  -7.048 1.82e-12 ***
JobSatisfaction                     -0.39157    0.05426  -7.216 5.34e-13 ***
WorkLifeBalance                     -0.19417    0.05367  -3.618 0.000297 ***
Age                                 -0.29562    0.07624  -3.878 0.000105 ***
NumCompaniesWorked                   0.34384    0.05739   5.991 2.08e-09 ***
TotalWorkingYears                   -0.52293    0.10418  -5.019 5.19e-07 ***
TrainingTimesLastYear               -0.23977    0.05605  -4.278 1.89e-05 ***
YearsSinceLastPromotion              0.54619    0.07626   7.162 7.96e-13 ***
YearsWithCurrManager                -0.57189    0.08722  -6.557 5.51e-11 ***
emp_outtime_int_avg                  0.58386    0.05289  11.040  < 2e-16 ***
`BusinessTravel - Non-Travel`       -0.84289    0.24274  -3.472 0.000516 ***
`BusinessTravel - Travel_Frequently`  0.83670    0.12637   6.621 3.57e-11 ***
`Department - Human Resources`       1.05609    0.21003   5.028 4.95e-07 ***
`JobRole - Research Director`        0.71374    0.22186   3.217 0.001295 **
`JobRole - Research Scientist`       0.49858    0.14113   3.533 0.000411 ***
`JobRole - Sales Executive`          0.46404    0.13769   3.370 0.000751 ***
`MaritalStatus - Divorced`          -1.06273    0.16093  -6.603 4.02e-11 ***
`MaritalStatus - Married`           -0.93152    0.12022  -7.749 9.28e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2865.5  on 3224  degrees of freedom
Residual deviance: 2227.8  on 3206  degrees of freedom
AIC: 2265.8
```

From this set of variables, management now can see which factors are the most important regarding their employees attrition.

3. Threshold optimization

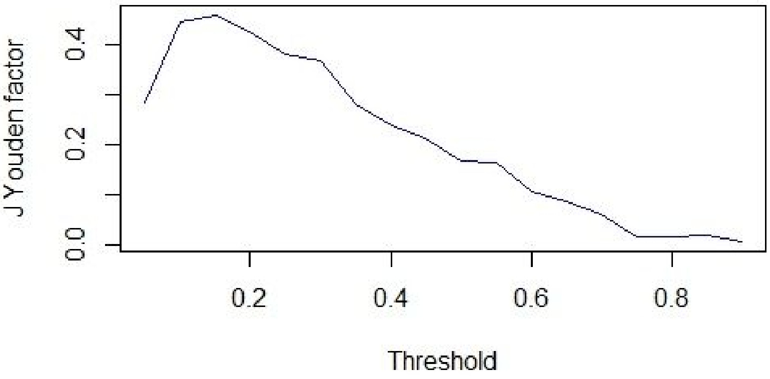
As we want to predict occurrence of attrition with the best possible precision we will not measure the quality of a model with Accuracy metric, but with Sensitivity metric which shows how precise model predict “1” - event occurrence.
For this kind of requests it is very convenient to use J Youden factor which is calculated as shown below, and visualized on Pic. 2.:

J = True Positive Rate - True Negative Rate +1



Pic. 2: J Youden factor visualization.

Attrition was predicted in a loop with changing threshold value from 0.05 to 0.9 with step equal to 0.05. Pic. 3. Shows results of threshold optimization using J Youden factor.



Pic. 3 Logistic Regression threshold optimization using J Youden factor

4. **Final thoughts**

Final model prediction results are shown in confusion matrix below:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 629 276
         1  40 130

               Accuracy : 0.706
                 95% CI : (0.6778, 0.7331)
    No Information Rate : 0.6223
    P-Value [Acc > NIR] : 5.103e-09

                  Kappa : 0.294
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.3202
            Specificity : 0.9402
         Pos Pred Value : 0.7647
         Neg Pred Value : 0.6950
             Prevalence : 0.3777
         Detection Rate : 0.1209
   Detection Prevalence : 0.1581
      Balanced Accuracy : 0.6302

       'Positive' Class : 1
```

As we can see Specificity and TPR have high value, model predicts event occurrence with high precision. What can bother is quite low overall model accuracy which equals 0.7.

Further work might be focused on improving model accuracy with maintaining high value of Specificity and TPR, or testing other methods.