SevvalTASYONAN

SEVVAL TASYONAN

LIBRARIES

We use caret package to train model and we use dalex package to roc curve.

```
library(caret)
library(DALEX)
```

1-) Description of the task

We have the data on customer's and their demographics. With this data, we are trying to find out whether it is risky to give credit to the customer. The data we have is encoded. We don't information about the features. Our target variable is **label** . It has two values 1 is high credit risk and 0 is low credit risk.

2-) Description of data set

We deleted the Id column because it didn't give us any information. Category features are fea_1, fea_3, fea_5, fea_6, fea_7, fea_9. So we transform them to factor. Other features are numerical. Data set is 1125 observation of 11 features and one target. Since we had enough number of observations, we removed the missing data from our dataset.

```
data <- read.csv("customer_data.csv")
data <- data[, -2]
data$label <- as.factor(data$label)
data$fea_1 <- as.factor(data$fea_1)
data$fea_3 <- as.factor(data$fea_3)
data$fea_5 <- as.factor(data$fea_5)
data$fea_6 <- as.factor(data$fea_6)
```

```r
data$fea_7 <- as.factor(data$fea_7)
data$fea_9 <- as.factor(data$fea_9)

data <- na.omit(data)
head(data)
```

```
  label fea_1  fea_2 fea_3  fea_4 fea_5 fea_6 fea_7 fea_8 fea_9 fea_10   fea_11
1     1     5 1245.5     3  77000     2    15     5   109     5 151300 244.9490
2     0     4 1277.0     1 113000     2     8    -1   100     3 341759 207.1738
3     0     7 1298.0     1 110000     2    11    -1   101     5  72001   1.0000
4     1     7 1335.5     1 151000     2    11     5   110     3  60084   1.0000
6     0     6 1217.0     3  56000     2     6    -1   100     3  60091   1.0000
7     1     4 1304.0     3  35000     2     8     9    85     5  60069   1.0000
```

3-) Train a logistic regression model.

We use train function from the caret the package to train logistic regression model. We use 10-fold cross validation to measure the performance of the model.

```r
set.seed(123)
control <- trainControl(method = "cv",
                        number = 10)
model <- train( label ~ .,
                data = data,
                trControl = control,
                method = "glm")
model
```

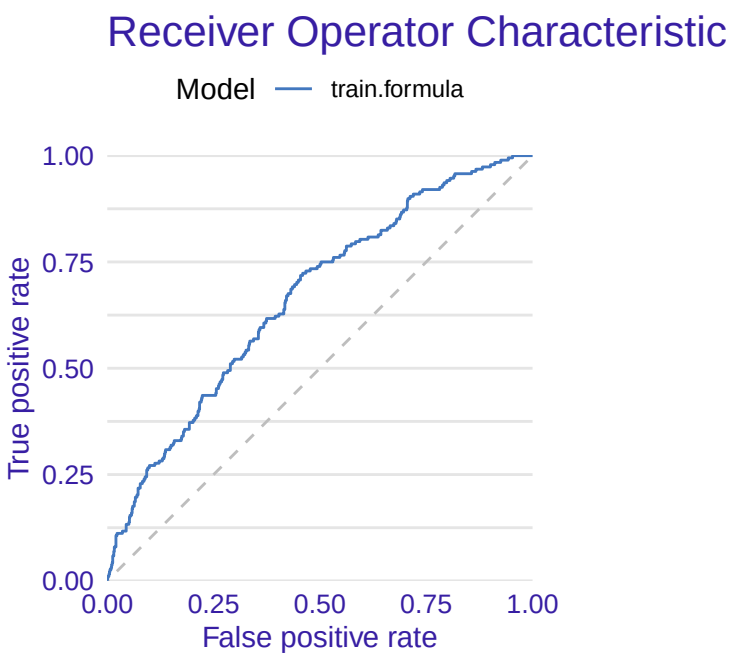
```
Generalized Linear Model

976 samples
 11 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 878, 878, 878, 878, 879, 879, ...
Resampling results:

  Accuracy   Kappa
```

```
0.8012159  -0.01170562
```

4-) Performance of the model

We use Roc Curve the major the performance.

```
explain_lr <- explain(model   = model,
                      data    = data[, -1],
                      y       = data$label == 1,
                      type    = "classification",
                      verbose = FALSE)
performance_lr <- model_performance(explain_lr)
plot(performance_lr, geom = "roc")
```



Roc Curve looks not bad but not very good. It is better than predicting randomly because it is above the line.

```
performance_lr
```

```
Measures for:  classification
```

```
recall     : 0.005319149
precision  : 0.5
f1         : 0.01052632
accuracy   : 0.807377
auc        : 0.6647687

Residuals:
           0%           10%           20%           30%           40%
-6.101237e-01 -2.816517e-01 -2.374070e-01 -2.026428e-01 -1.685216e-01
          50%           60%           70%           80%           90%
-1.430284e-01 -1.176322e-01 -8.030589e-02 -1.199214e-12  7.698937e-01
         100%
 9.896181e-01
```

Auc is 0.66 acceptable.

5-) Imbalance problem

We looked at imbalance problem. Since the values of 0 and 1 are not close to each other, there may be an imbalance problem.

```r
summary(data$label)
```

```
  0   1
788 188
```

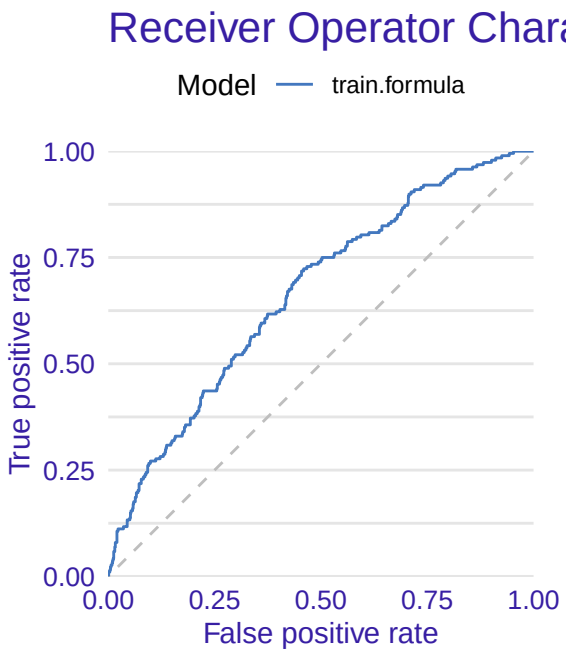
We use stratified cross validation because there maybe imbalance

```r
set.seed(123)
folds <- 10
cvIndex <- createFolds(factor(data$label), folds, returnTrain = T)
tc <- trainControl(index = cvIndex,
                   method = 'cv',
                   number = folds)

model2 <- train(label ~ ., data = data,
                method = "glm",
                trControl = tc)
model2
```

```
Generalized Linear Model

976 samples
 11 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 879, 878, 879, 879, 879, 878, ...
Resampling results:

  Accuracy   Kappa
  0.8063539  -0.001997897
```

Two models have equal accuracy so we think there is no imbalance problem.

```r
explain_lr <- explain(model   = model2,
                      data    = data[, -1],
                      y       = data$label == 1,
                      type    = "classification",
                      verbose = FALSE)
performance_lr <- model_performance(explain_lr)
plot(performance_lr, geom = "roc")
```




```
performance_lr
```

```
Measures for:  classification
recall     : 0.005319149
precision  : 0.5
f1         : 0.01052632
accuracy   : 0.807377
auc        : 0.6647687

Residuals:
           0%           10%           20%           30%           40%
-6.101237e-01 -2.816517e-01 -2.374070e-01 -2.026428e-01 -1.685216e-01
          50%           60%           70%           80%           90%
-1.430284e-01 -1.176322e-01 -8.030589e-02 -1.199214e-12  7.698937e-01
         100%
 9.896181e-01
```

Auc also equal between the model1 and model2