# EARIN

# MODEL PROJECT

SEBASTIAN GRZELAK

## Data description:

The file “WorkersInCompany.csv” contains the 15000 data set with a given columns:

1) **satisfaction_level**
   - value from interval [0;1] which represents the satisfaction level of given worker
2) **last_evaluation**
   - value from interval [0;1] gives information about last worker evaluation
3) **number_project**
   - number of project in which worker take part
4) **average_montly_hours**
   - average time of work in company in hours
5) **time_spend_company**
   - extra time spend in company
6) **work_accident**
   - number of accident for given worker
7) **left**
   - nalue 0 -> worker wants to stay in company, 1 -> worker wants to live comapany
8) **promotion_last_5years**
   - number of promotion in last 5 years
9) **department**
   - name of the worker department
10) **salary**
   - low, medium, high those values represent the worker salary

## Data selection:

Columns which were selected in my model:

- **satisfaction_level**
- **average_montly_hours**
- **promotion_last_5years**
- **salary**
- **time_spend_company**

Those columns have the highest influance on the final accurancy and prediction of the model.

## Final results:

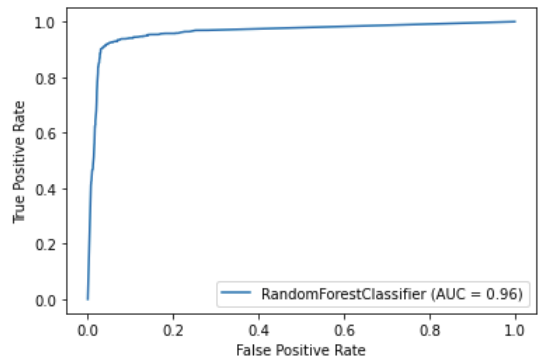
## 1) Random Forest Classifier Model

```
from sklearn.metrics import accuracy_score, precision_score, recall_score

acc = accuracy_score(y_test, y_pred)
pre = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)

print('Accuracy:', acc)
print('Precision:' , pre)
print('Recall:', recall)
```

```
Accuracy: 0.9518333333333333
Precision: 0.8966346153846154
Recall: 0.9040858725761773
```

```
plot_roc_curve(model, X_test, y_test)
```

```
<sklearn.metrics._plot.roc_curve.RocCurveDisplay at 0x1baef890508>
```



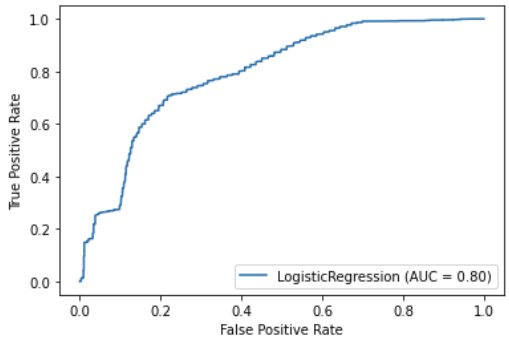
## 2) Logistic Regression Model

```
acc = accuracy_score(y_test, y_pred)
pre = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)

print('Accuracy:', acc)
print('Precision:' , pre)
print('Recall:', recall)
```

```
Accuracy: 0.7740833333333333
Precision: 0.5653136531365314
Recall: 0.26523545706371193
```

```
plot_roc_curve(model, X_test, y_test)
```

```
<sklearn.metrics._plot.roc_curve.RocCurveDisplay at 0x1baf057f448>
```




As can be observed the **Random Forest Classifier** model gave the results with much better precision and accuracy than the **Logistic Regression** model. The difference is around 20% in case o accuracy and more than 30% in case of precision. Such a diffrence has a huge impact on correct and efficient decisions by choosen model.