University of Siegen Faculty III: School of Economic
Master Program in Economic Policy
January 30, 2022

Seminar Project

Which employee will leave the company next?

Seminar: Solving Big Data Problems

Examiner: Dr. Alexander Hoffmann
Author: Daria Kharitonova

Siegen

Content

Business problem

- Corporation X with 14999 data on employees (anonymous data)
- Why are **the best** and the **most experienced employees leaving prematurely**?
- TASK: How to predict that an employee leaves the company before it happens?
- SOLUTION: Collect the data on employees and apply ML algorithm

- In ML framework this is Binary Classification task: 2 groups of employees

→ **Leave vs. Don't Leave**

Variables Description

Variable Name	Description	Type of Variable
satisfaction_level	Satisfaction	Numeric, continuous
last_evaluation	Last review	Numeric, continuous
number_project	Number of projects done by employees	Numeric, discreate
average_montly_hours	Average working hours per month	Numeric, discreate
time_spend_company	Years from entry time	Numeric, discreate
work_accident	Whether there is a work accident	Binary -> Dummy
promotion_last_5years	Have you been promoted in the last five years	Binary -> Dummy
department	Staff department	Categorical -> Dummy
salary	Salary level	Categorical -> Dummy
left	**Resign**	**Categorical -> Dummy**

Data investigation and analysis

- There are 14999 observations in 10 columns with no missing values

```
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   satisfaction_level     14999 non-null  float64
 1   last_evaluation        14999 non-null  float64
 2   number_project         14999 non-null  int64  
 3   average_monthly_hours  14999 non-null  int64  
 4   time_spend_company     14999 non-null  int64  
 5   Work_accident          14999 non-null  int64  
 6   left                   14999 non-null  int64  
 7   promotion_last_5years  14999 non-null  int64  
 8   department             14999 non-null  object 
 9   salary                 14999 non-null  object 
dtypes: float64(2), int64(6), object(2)
```

Data investigation and analysis

Descriptive Statistics

Variables	mean	std	min	max
satisfaction_level	0.61	0.25	0.09	1
last_evaluation	0.72	0.17	0.36	1
number_project	3.80	1.23	2	7
average_monthly_hours	201.05	49.94	96	310
time_spend_company	3.50	1.46	2	10
Work_accident	0.14	0.35	0	1
left	0.24	0.43	0	1
promotion_last_5years	0.02	0.14	0	1

+ Investigation of distribution plots and histograms*

Key insights

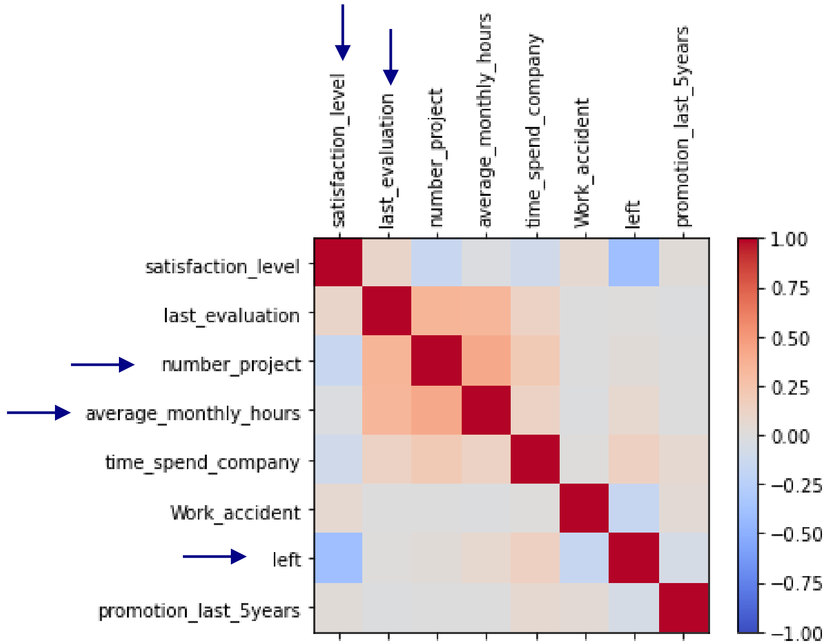
- Despite the fact that 55% of employees have level satisfaction (> 0.6), 13% have the level of satisfaction (<0.3)
- Working extra hours was a norm in the company: on average employees do ~ 16 extra hours per month (if norm = 184 hours = 23 full days per month). 59% of all employees work more than 184 hours per month, 31% - more than 201 hours per month. It is unclear if those extra hours are paid
- On average an employee is assigned to 3-4 projects, but at least to 2
- Level of promotion is extremely low: 2%
- After 6 years the probability that an employee leaves the company is low
- 49% get low salary and 43% medium salary. Only 8% get high salary. However, thresholds of the salary are not given.
- 41% of all employees work in Technical, Support and IT departments, 28% in sales

* PS. Distribution graphs and histograms of the key variable can be found in the back-up slides

Data investigation and analysis

UNIVERSITÄT SIEGEN

CORRELATION MATRIX



Key insights:

- Strong negative correlation of satisfaction level, number of projects and decision to leave. → The more employees were working and the lower was the satisfaction level, more often they decided to leave the company
- Positive correlation of high last evaluation and number of projects and average monthly hours → The more employees were working, the higher they were evaluated

Empirical framework

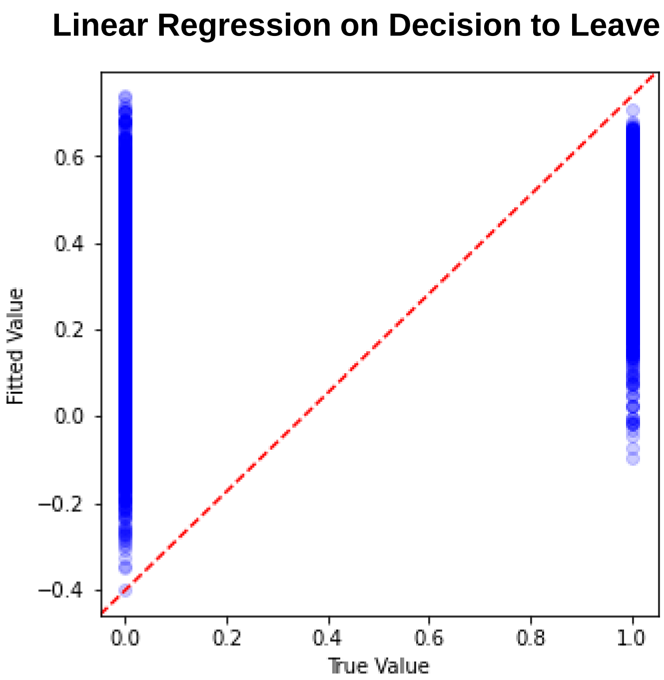
- **Predicted variable:** Y = df.left → decision to leave the company, either 1 or 0
- **Predictable variables:** X = df.drop(['left'], axis = 1) → all other variables
- **Split of the data on test and train:** X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=.30, stratify=y, random_state=101) → 10499 for Train set and 4500 for Test set
- **Standardize the variables**
- **Initiate the model →** LinearRegression(), LogisticRegression(), DecisionTreeClassifier(), RandomForestClassifier(), XGBClassifier()
- **Fit the model** → Make predictions → Calculate Accuracy, Precision, Recall, and F1 and AUC scores → Display Confusion Matrix
- Perform **Hyper Parameter Tuning:** GridSearchCV and RandomizedSearchCV
- **Choose the best model:**
  - **Situation:** The company wants to find a balance between Precision and Recall, resources are limited → **best AUC score**

* X = df.drop(['left', 'average_monthly_hours'], axis = 1) for Linear Regression to address multicollinearity

Why linear regression is not suitable?

- **Problem #1: Predicted value is continuous, not probabilistic**
- **Problem #2: Sensitive to imbalance data:**
  - 0 11428
  - 1 3571
- Fit the model to the train dataset
- Predict Y on the test dataset
- Look at the **key metrics**:
  - → $R^2 = 0.205, \text{MSE} = 0.1439$
- **→ Linear Regression is not the best fit for the binary classification task**

True vs. Fitted Values

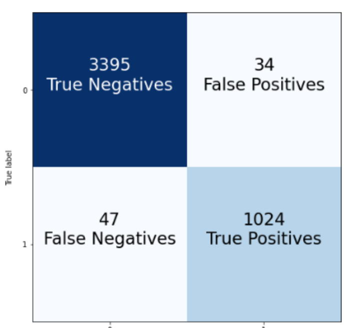


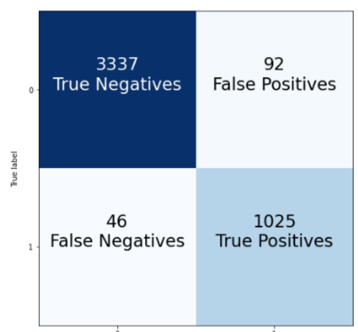
Models Comparison

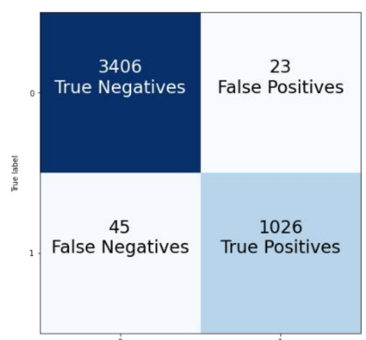
Models' Scores: Selection Criteria – AUC Score

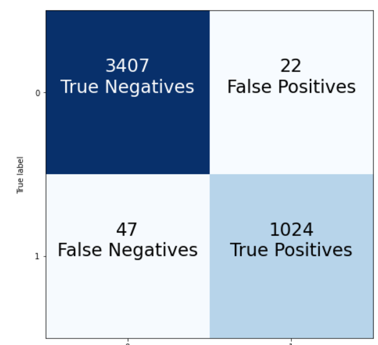
	Logistic Regression	Decision Tree	Random Forest	XGBoost
Default Model	Accuracy : 0.792 Precision: 0.604 Recall : 0.366 F1 score : 0.456 **AUC score: 0.822**	Accuracy : 0.969 Precision: 0.918 Recall : 0.957 F1 score : 0.937 AUC score: 0.965	Accuracy : 0.985 Precision: 0.980 Recall : 0.957 F1 score : 0.968 **AUC score: 0.991**	Accuracy : 0.981 Precision: 0.970 Recall : 0.949 F1 score : 0.959 AUC score: 0.989
Tuned model	Accuracy : 0.802 Precision: 0.729 Recall : 0.264 F1 score : 0.388 AUC score: 0.764	Accuracy : 0.975 Precision: 0.965 Recall : 0.928 F1 score : 0.946 **AUC score: 0.974**	Accuracy : 0.983 Precision: 0.977 Recall : 0.952 F1 score : 0.965 AUC score: 0.990	Accuracy : 0.983 Precision: 0.972 Recall : 0.955 F1 score : 0.963 **AUC score: 0.991**

Confusion Matrix of the best model





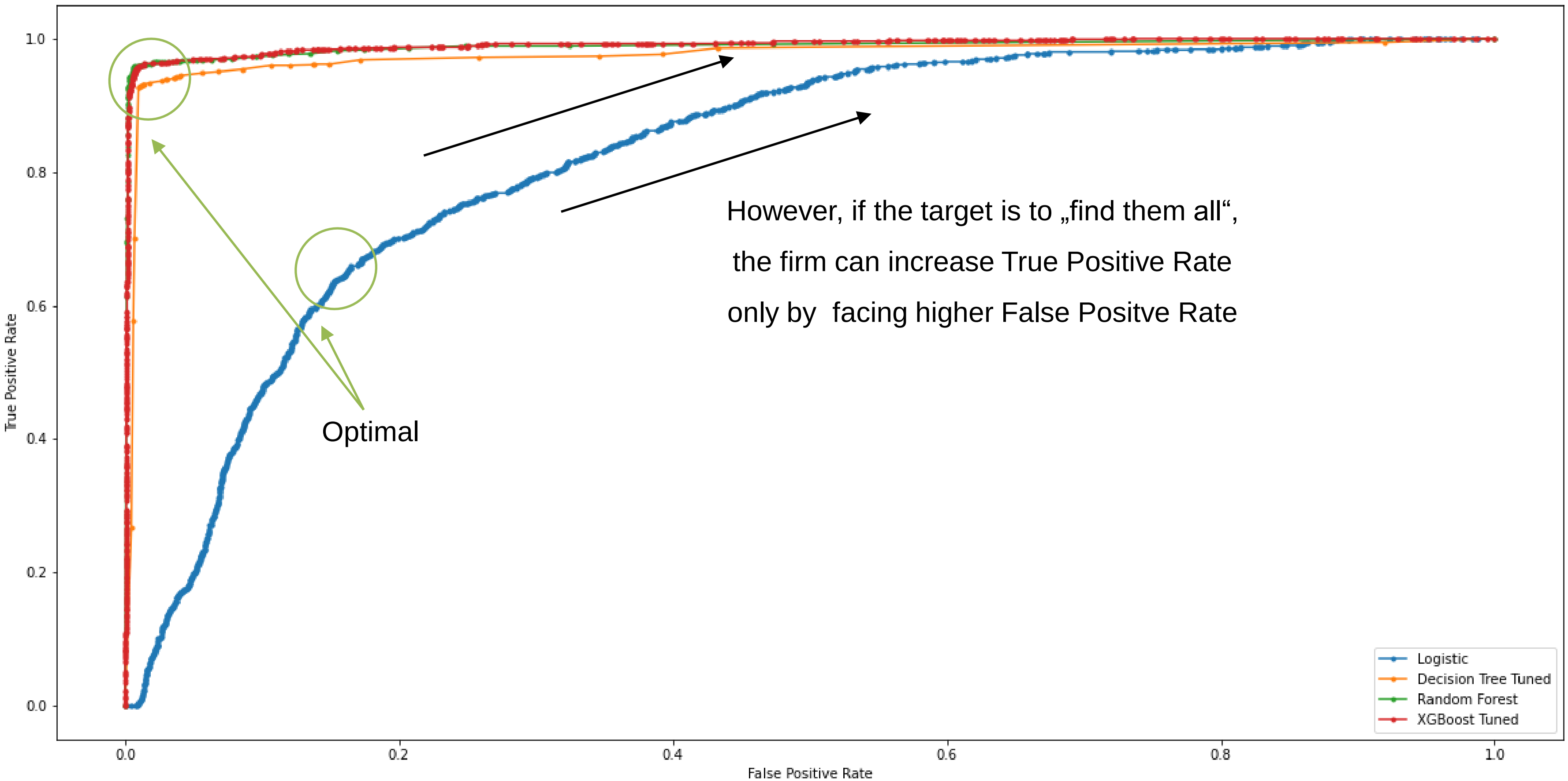




→ **Random Forest and XGBoost produce the highest AUC score, however, other parameters in Random Forest are slightly higher**

Models Comparison

AUC Curves for the Models with the best parameters



- Best models: Random Forest, XGBoost hyper-parameter tuned

Conclusions

- Applying ML the company can identify up to 96% of all employees who are most likely leave the company and undertake particular measures to avoid this scenario
- In the context of this company the algorithms with the highest performance are Random Forest and XGBoost
- If the company wants to be able to identify all employees who are going to leave the company, it has face high False Positive Rate
- The company can also undertake preventive measures and change its policies in order to reduce staff turnover such as, e.g. ensuring healthy work-life-balance and avoiding to evaluate employees with higher extra hours higher than others
- If any major changes in human resources policy of the company will take place, it is recommended to evaluate model once again

Thank you!

Back- up slides

Data investigation and analysis

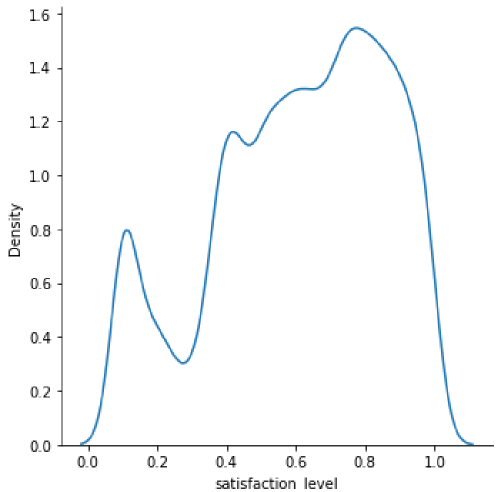
UNIVERSITÄT SIEGEN

DESCRIPTIVE STATISTICS

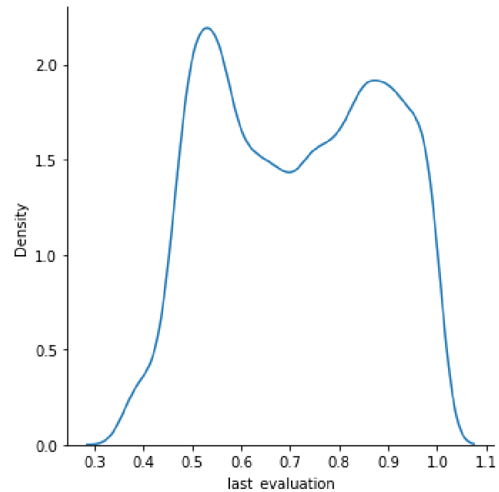
Variables	mean	std	min	max
satisfaction_level	0.61	0.25	0.09	1
last_evaluation	0.72	0.17	0.36	1
number_project	3.80	1.23	2	7
average_monthly_hours	201.05	49.94	96	310
time_spend_company	3.50	1.46	2	10
Work_accident	0.14	0.35	0	1
left	0.24	0.43	0	1
promotion_last_5years	0.02	0.14	0	1

KEY VARIABLES VIZUALIZATION

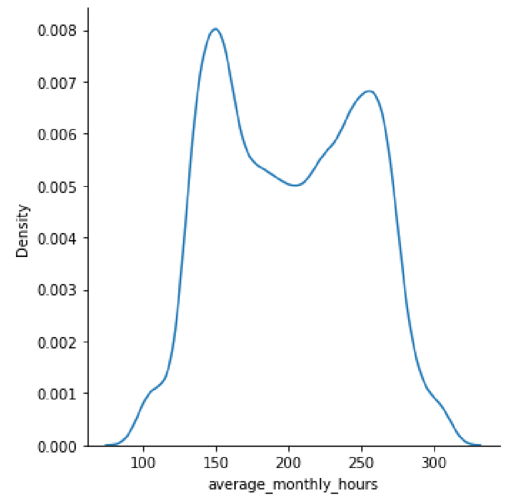
Satisfaction Level



Last Evaluation Score



Av. Monthly hours



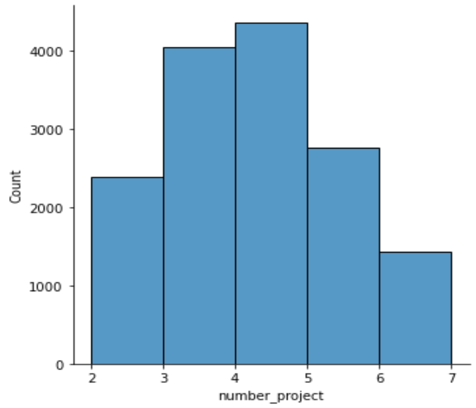
Data investigation and analysis

DESCRIPTIVE STATISTICS

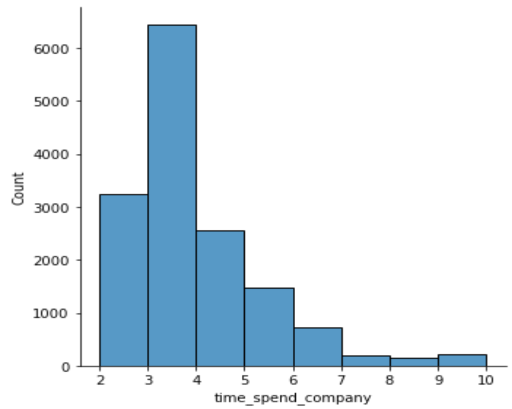
Variables	mean	std	min	max
satisfaction_level	0.61	0.25	0.09	1
last_evaluation	0.72	0.17	0.36	1
number_project	3.80	1.23	2	7
average_monthly_hours	201.05	49.94	96	310
time_spend_company	3.50	1.46	2	10
Work_accident	0.14	0.35	0	1
left	0.24	0.43	0	1
promotion_last_5years	0.02	0.14	0	1

KEY VARIABLES VIZUALIZATION

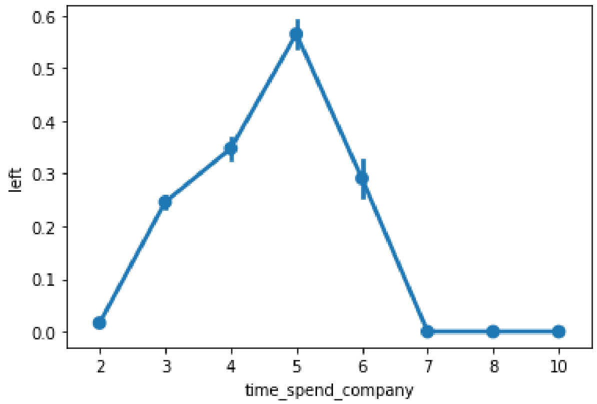
Number of Projects



Time Spent in the Company: Total



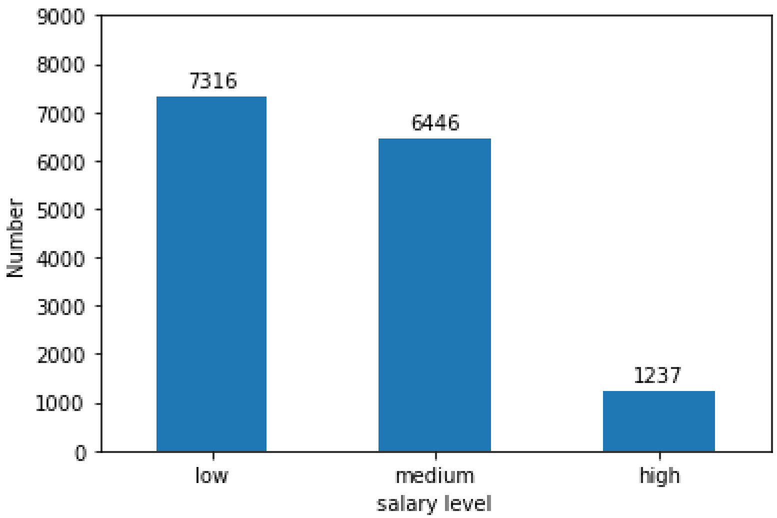
Time Spent in the Company: Left = 1



Data investigation and analysis

KEY VARIABLES VIZUALIZATION

Number of Employees by Salary Level



Number of Employees by Department

sales	4140
technical	2720
support	2229
IT	1227
product_mng	902
marketing	858
RandD	787
accounting	767
hr	739
management	630