|  |  |
| --- | --- |
| Presentation date: | 21.06.2024 |
| Topic: | **Machine Learning Algorithms to Predict Employee Turnover** |
| Course: | Machine Learning |
| Lecturer: | Elizabeth Staegemann |
| Data source: | Kaggle  [Employee Turnover (kaggle.com)](https://www.kaggle.com/datasets/davinwijaya/employee-turnover/data) |
| Data type used: | Csv, ipynb, python file |
| Presenters: | Qin, Jinwen  Hrioua, Jad  Serrano Redondo, Antonio  Madjet, Mohamed El-Amine  Mukhitdinova, Lina |

**Table of contents:**

1. Introduction
2. Goal of the project
3. Data preparation
4. Describing the data

**Introduktion:**

According to a study, losing a single employee can cost tens of thousands of dollars, meaning a wave of employee turnover can quickly escalate costs into the hundreds of thousands. Additionally, a newly hired employee may take one to two years to reach the productivity level of an existing employee. Therefore, companies can’t afford such a financial or productivity hit. Companies spend money and time recruiting talented people, and if some of them decide to leave, it represents a loss of investment. Consequently, companies can save money if they can intervene before their employees leave. In fact, measuring employee turnover can be helpful to employers who want to examine reasons for turnover or estimate the cost-to-hire for budgeting purposes.

**Client:** (Auftragsgeber): The HR department wants to know and understand why employees are resigning. They collected data from employees, but now they don’t know what to do with it. They asked us to provide data-driven suggestions based on our understanding of the data.

They have the following question: What’s likely to make the employee leave the company?

The goals in this project are to analyze the data collected by the HR department and to build a model that predicts at 80 % whether or not an employee will leave the company.

If we can predict employees likely to quit, it might be possible to identify factors that contribute to their leaving.

**Goal of the project**: By trying different machine learning models, we want to find out which model best predict whether an employee leaves or not and determine the factors which affect his decision making.

**Hypothesis:** *Why are people resigning?*

The most common reasons for resigning from a previous job are:

* Lack of career development/advancement
* Inadequate compensation
* Lack of meaningful work
* Lack of work flexibility
* Employees looking for better opportunities
* Bad management.
* Excessive working hours.
* No promotion
* Not satisfied with the current job
* A negative working environment.

Our objective is to get a score around 80 %, for this we tried to use the following machine learning algorithms starting with supervised ones, as our study deals with classification:

* Bayes (**Lina**)
* Logistic Regression (**Mohamed**)
* Random Forest (**Jad**)
* K-Nearest Neighbors algorithm (KNN) (**Jinwen**)
* Support Vector Machine (SVM) (**Jinwen**)
* Multilayer Perceptron (MLP) (**Jinwen**)
* Decision Tree (**Antonio**)

**Investigation Steps:**

* **Load the corresponding dataset** 
  + Analyze the dataset (looking for outliers, checking missing values, which columns are categorial, and which are numeric, which are feature and which are target).

Since our dataset contains categorical variables, we need to convert them into numerical representations that can be understood by the model.

-Which columns are not relevant to the current study or can’t provide useful information.

-We checked also the dependence between different features.

-Check which categories can be combined (see mapping.xls)

-We considered all the columns, hoping that the algorithms can find useful features.

**Data preparation**

1. First, pandas is used to load the ‘turnover.csv’ dataset: df = d.read\_csv('turnover.csv', encoding = 'ISO-8859-1');
2. The dataset has 1129 rows (employees) and 16 columns, no missing values was detected; The dataset looks clean and there are no missing values;
3. df.describe() provides also mean values and for ‘event’ one gets 50.6 %.

Balanced dataset: 50.6 % of the employees left the organization and 49.4 % did not leave the organization making our dataset to be considered as balanced.

1. Categories combining and Encoding

For the columns industry, profession, coach and traffic we combined categories.

We also used two encoders:

One Hot encoder: industry, profession, traffic, coach

Label encoder: way

The rest of the features were already encoded.

The new dataset is saved in a new file: 'df2\_encoded.csv', which is used for different machine learning algorithms.

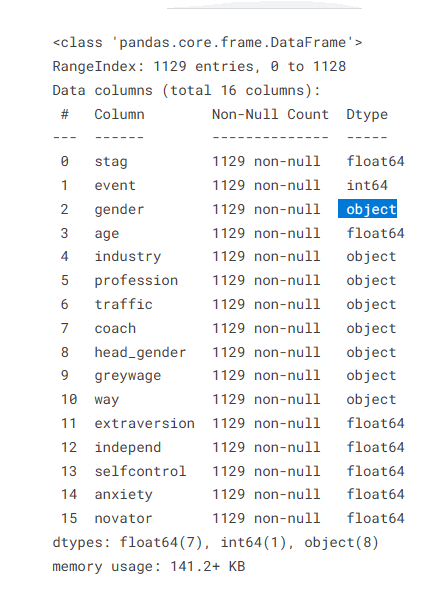
From this step, each of us use this dataset and test the corresponding machine learning model.

1. Scaling We used two approaches: first without scaling the data and the second approach is scaling the data with both StandardScaler and MinMaxScaler.

**Describing the data**

**Columns Attributes**

We have the following attributes column-wise:



* stag - Experience (time) in months
* event - Employee turnover, this hast two possible values: 1 for resigning and 0 for staying
* gender - Employee's gender, female(f), or male(m)
* age - Employee's age (year)
* industry - Employee's Industry
* profession - Employee's profession
* traffic: From what pipeline employee came to the company:

1. KA: The recruiting agency brought you to the employer.
2. Advert: contacted the company directly, after learning from advertising.
3. Empjs: The employer contacted the candidate after seeing his resume on the job site.
4. Friends: The employer contacted you on the recommendation of a person who knows you
5. rabrecNErab: the employer contacted a candidate on the recommendation of a person who knows him.
6. recNErab: contacted the company directly on the recommendation of your friend - NOT an employee of this company.
7. Referal: contacted the company directly on the recommendation of your friend - an employee of this company
8. youjs: You have applied for a vacancy on the job site

* **Coach**: Presence of a coach (training) during probation. This categorical attribute has the following values: (Yes/ No/ my head)
* head\_gender: head (supervisor) gender: male or female.
* greywage: The salary does not seem to the tax authorities

white: the employee becomes only salary,

grey: the employee gets salary and other benefits (car, phone..)

* way: Employee's way of transportation.

1. Bus: takes the bus to go to work.
2. Car: the employee drives to work with his car.
3. Foot: the employee goes by foot to work.

* **Extraversion**: personality trait characterized by excitability, sociability, talkativeness, assertiveness, and high amounts of emotional expressiveness.
* **Independent**: the ability to make and carry out important decisions, without outside coaching.
* **Selfcontrol**: The ability to control behaviors to avoid temptations and achieve goals.
* **Anxiety**: People who tend to experience apprehension, tension, or uneasiness that stems from the anticipation of danger, which may be internal or external.
* **Novator**: An innovator.

We differentiate between the following attributes of the dataset and their type

**Categorial**: gender, industry, profession, traffic, coach, head, gender, greywage, way

**numerical**: stag, event, age, extraversion, independent, selfcontrol, anxiety, novator

**Observations/ hypotheses about the dataset:**

**Age**

**Category age,** taking a subset of the age range:

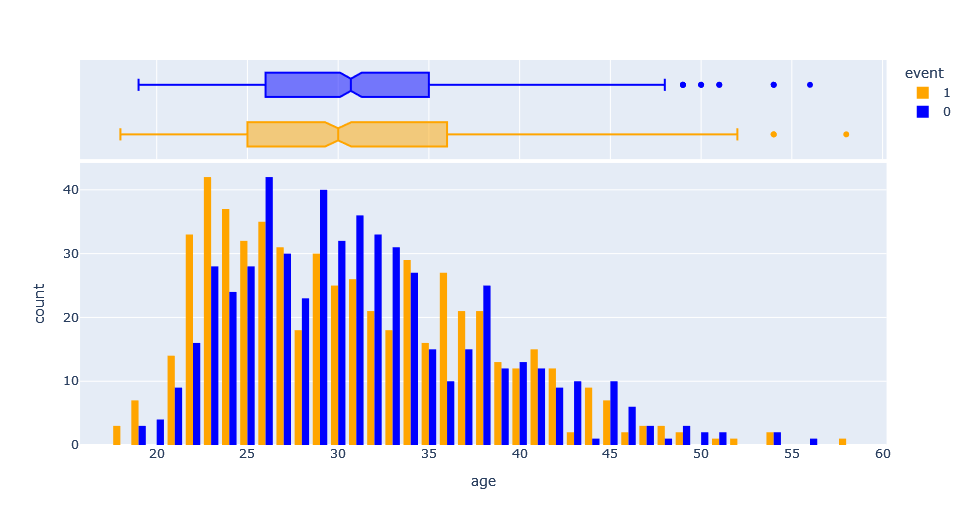
* **first category**: from 18 to 25 years.

the reason why employees from this category are resigning is due to unsatisfaction with the Chef/Coach or the employees are not happy with the choice they made for the training/Ausbildung.

* **the second one** is from 26 to 36 years:

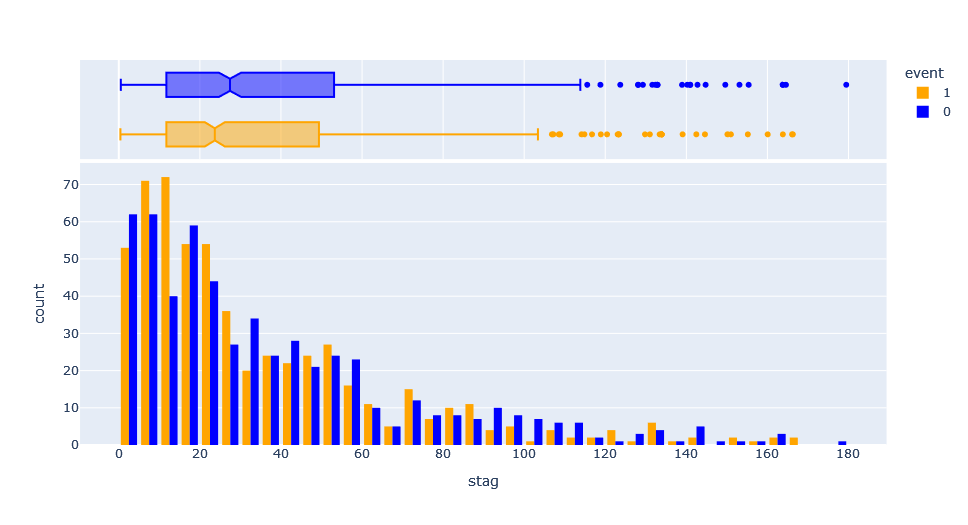
most of the employees are staying as they want probably to gain more expertise and hope for promotion.

* **the third range** is from 37 to and 45 years: half decided to stay, and half left the company. Maybe the employees who left are looking for more perspectives and more challenges.
* **the fourth one** is for those older than 45 years, at this age, the employees are looking for stability and they prefer to stay rather than leaving the company as the chances of getting a new job is very limited.



**Stag –** describes the time spent in the company and it is given in months

* **From 0 to 5 moth**: probation period.
* **From 6 to 30 months**: fixed term contract, not happy with the new job.
* **From 31 to 59 months:** after gaining expertise, employees are leaving after looking for a better paid job and with more challenges.
* **From 60 to 90 months**: after 5 years of work in a company, some employees who are not getting promoted tend to resign.
* **Between 90 and 120 months:** after more than 7 years of employment, most employees prefer to stay; seeking stability.



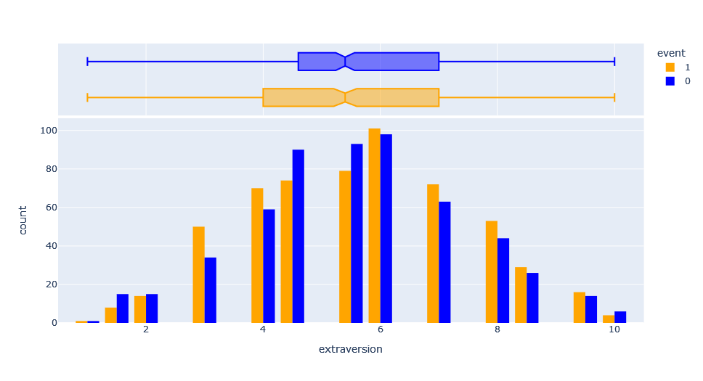
**In our opinion: What is missing from the dataset to identify the main reason why the employees are leaving their job:**

* Salary amount & Working number of hours per week -> could contribute to burnout;
* Position: senior, junior, manager or trainee;
* Business trip/ hybrid or Working from home possibilities;
* Possibility for promotion or whether the employee was promoted in the past;

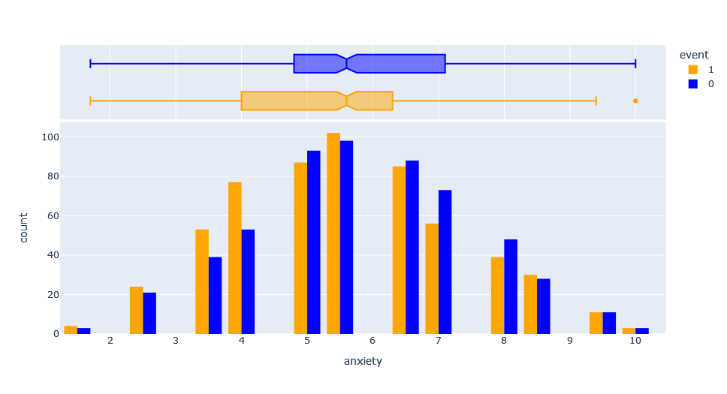
**Selfcontrol, Novator and Independent features seem to mimic the Normal distribution:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | |
| A graph with different colored bars  Description automatically generated | |  | |

**Extraversion** – for extraversion value less than 4.6, more employees left compared to those who decided to stay.

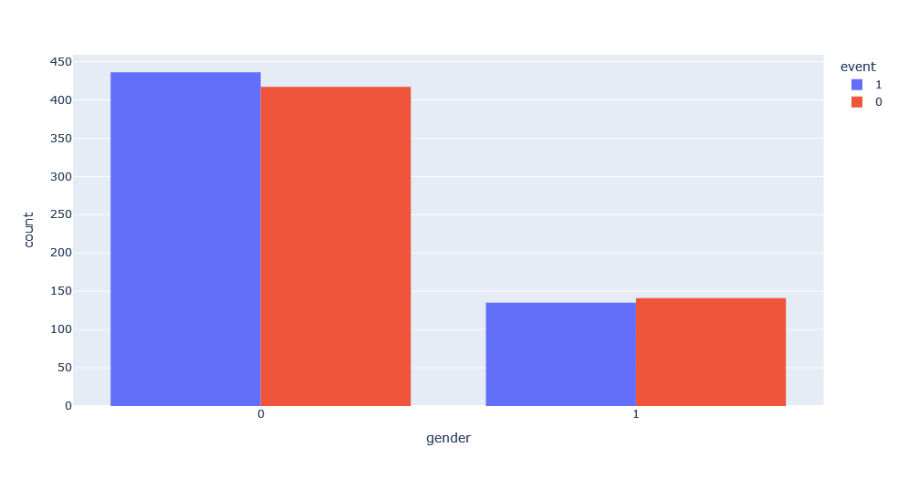


**Anxiety** - for an anxiety value between 0 and 4.8, more employees decided to leave, taking the risk to look for a new job. Then, for a value ranging from 4.8 to 6.3, half of the employees resigned. For a value greater than 6.3, employees prefer stability as they are scared of not finding a new job if they decide to leave the company, feeling they are in a comfort zone. This shows a direct correlation between leaving/staying and anxiety, and therefore, this column is important to consider in our study.



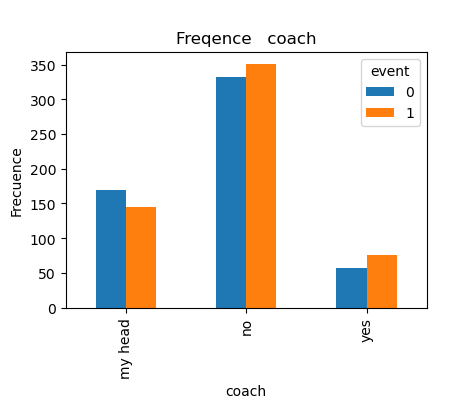
Now, let’s extend our investigation to other columns:

**Gende**r: As we can see on the following figure, that almost 50% of man and woman resigned and the same percent stayed.

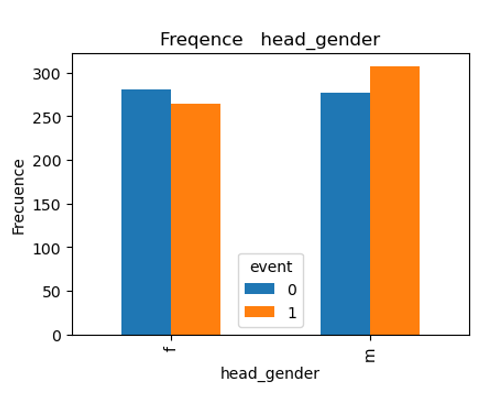


By looking at this graph we don’t get additional information. However, we should consider it in our analysis as the other features could be linked to this one.

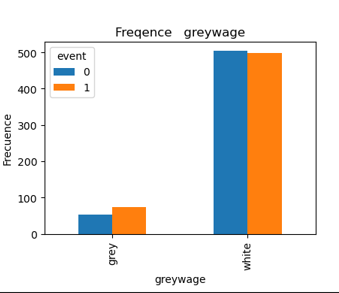
**Coach:** We can reduce these one to two possibilities, yes (by combining my head and yes) and no.



**Head gender:**



**Grey wage:**



**Intensive Testing every Model**

**For all the models the data was divided into Training and Testing sets:**

It is crucial to evaluate our models’s performance on unseen data to assess their ability to generalize. We use the train\_test\_split() function to split the data into training and testing sets. The X\_train, X\_test, y\_train, and y\_test variables store the resulting split, with 80% of the data used for training and 20% for testing.

Tr=Training, te=testing

show

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | BGs | DT | RF | LR | KNN | SVC | MLP |
| trainingsscore | 0.58(tr) | 1 (tr) | 0.65 (te)  1 (tr) | 0.59(tr) | 0.74(tr) | 0.58 | 0.76 |
| crossvalscore | 0.53(tr) | 0.61 (tr) | 0.63 (tr) | 0.56(tr) | 0.58(tr) | 0.55 | 0.58 |
| Grid-SearchCV | 0.58(tr) |  | 0.64 | 0.59(tr) | 0.64(tr) | 0.55 | 0.59 |
| Score auf Testmenge | 0.53 |  | 0.588t(te) | 0.56(te) | 0.66(te) | 0.54 | 0.55 |

**Fine Tuning Model**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
| RandomForest(Jad) | * Fit with default: **MinMax Scaling is best** * Cross\_Val\_Score :9 folder is best * GridSearchCV: {'bootstrap': False, 'criterion': 'entropy', 'max\_depth': None, 'max\_features': 6, 'min\_samples\_leaf': 2, 'min\_samples\_split': 5, 'n\_estimators': 30   }   * On Test Menge: 0.588 |
| LogReg(Mohamed)  + Interaction terms | * Fit with default: MinMax **Scaling is best** * Cross\_Val\_Score :3 folder is best * GridSearchCV: {'C': 1, 'penalty': 'l2', 'solver': 'saga'} * On Test Menge: 0.56 * On Test Menge: 0.62 |
| KNN(Jinwen) | * Fit with default: **no Scaling is best** * Cross\_Val\_Score :8 folder is best * GridSearchCV: {'algorithm': 'auto', 'leaf\_size': 10, 'metric': 'manhattan', 'n\_neighbors': 9, 'weights': 'distance'} * On Test Menge: 0.66 |
| SVC(Jinwen) | * Fit with default: **no Scaling is best** * Cross\_Val\_Score :7 folder is best * GridSearchCV: Best parameters: {'decision\_function\_shape': 'ovo', 'kernel': 'linear'} * On Test Menge: 0.54 |
| MLP(Jinwen) | * Fit with default: **Standard Scaling is best** * Cross\_Val\_Score :7 folder is best * GridSearchCV: Best parameters: {'activation': 'tanh', 'alpha': 0.0001, 'hidden\_layer\_sizes': (30,), 'learning\_rate': 'constant', 'solver': 'adam'} * On Test Menge: 0.55 |
|  |  |

# **Logistic Regression model**

### **Introduction**:

Classification is an area of supervised machine learning that tries to predict which class or category some entity belongs to, based on its features. Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. **This ML algorithm is a technique that models a categorical dependent variable (Y) based on one or more independent variables(X).** Logistic regression uses the concept of predictive modeling as regression; therefore, it is called logistic regression, but is used to classify samples; Therefore, it falls under the classification algorithm.Logistic regression predicts the output of a categorical dependent variable. Consequently, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

### **Logistic Function (Sigmoid Function)**

The sigmoid function is a mathematical function used to map the predicted values to probabilities. It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the **Sigmoid function** or the logistic function.

### **Assumptions for Logistic Regression**

1. The dependent variable must be categorical and binary.
2. There should be little or no multicollinearity between the independent variables.
3. There is linear relationship of variables to log odds.
4. Requires sufficiently large sample size.
5. There are no extreme outliers.
6. Should have independent observations.

### **Steps in our Logistic Regression model**:

1. Import packages, functions, and classes
2. Used non scaled data to compute the scores on training and test data, we got respectively, 0.58 and 0.56.
3. Used training Data with Standard scaling and 0.58 and 0.55 for training and testing data, respectively.
4. Used MinMax scaling and got 0.59 and 0.56, respectively for training and testing scores.
5. MixMax scaling seems to pride slightly better scores, therefore this scaling was used
6. For Cross Value score, we found 3 folds to lead to the best score
7. Using gridSearch we identified the best parameters for the model as: Hyperparameters: {'C': 1, 'penalty': 'l2', 'solver': 'saga'}, we obtained the following training and test score: 0.59 and 0.56, respectively. In general, using different options for the model had almost no pronounced effect on improving the scores. Maybe this has to do with the nature of our dataset.
8. We calculated the coefficients sorted by the magnitude of each feature's impact on employee turnover. In Absolut value, age, anxiety and stag have the largest coefficients.
9. Incorporated the interaction terms into the logistic regression model to assess whether this enhancement leads to improved model performance. We introduced all possible interactions, this extensive set of interactions could lead to overfitting, making the model too specific to the training data and potentially hindering its generalization to new, unseen data. Maybe by strategically selecting a subset of interactions, we can maintain model interpretability and efficiency while capturing the essential nuances in the relationship between different predictors and event (employee resigning). If we had more time we could investigate more this by selecting the interactions terms.

**Lessons Learned**

* The teamwork was a great experience, we had interesting discussions during the last three days, exchanged ideas and learned from each other.
* I gained experience from this project, and learned how to prepare the data in a format that is more appropriate for machine learning algorithms. I learned how to use scaled and non-scaled and scaled data and compute score on test and train data, how to improve the parameters of the ML model.
* The ‘event’ in the dataset, has a week correlation with other features which makes probably hard to get high scores. In fact, the dataset, we selected seems was not the best one, as it is missing many important features which could directly impact the employee resignation, like salary, working hours, satisfactions, promotion.
* In my opinion, we need more time to conduct a project on Machine Learning and get a deep understanding, and to tune correctly the different parameters of the model. For example, for the logistic regression model, it seems including the interaction terms lead to better score, but I did not get enough time to incorporate selective interaction terms.