# Data Analysis on the Factors that Affect Employee Turnover in a Company

**Weiyan XIE**
wxieai@connect.ust.hk

**Song MA**
smaal@connect.ust.hk

**Zihan ZHANG**
zzhangeo@connect.ust.hk

**Zhili WANG**
zwangeo@connect.ust.hk

**Xuantong LIU**
xliude@connect.ust.hk

## 1 BACKGROUND INFORMATION AND AIM

### 1.1 Background Information

Employee turnover is an important and unavoidable issue in many companies and organizations. Many experts of human resource management suggest that it is hard to access the real cost of losing employee because the leaving of employee will bring many other costs such as the recruitment cost, training cost of new employee and so on. Some studies (such as Society for Human Resource Management) predict that every time a business replaces a salaried employee, it costs 6 to 9 months' salary on average.

Therefore, one of the most important tasks for Human Resource Department in a company is to retain the valuable employee and keep a reasonable employee turnover rate. In this past, the implementation of such kind of tasks is time-consuming and mainly relies on the experience of human resource professionals. However, in recent years, the new technology such as artificial intelligence (AI), machine learning (ML), and deep learning (DL) is started to be applied in the field of human resource. These technology can help HR in terms of recruiting and retention, and reducing bias in HR decision-making. Some companies like IBM has already been beneficial from the application of these technology.

The above information inspires us to start this project to use the data analysis techniques in terms of employee retention in a company. The detailed aims of our project are given in the following part.

### 1.2 Aims of the project

From the angle of data analysis, the aim of this project is to apply the machine learning methods to find out how the employee turnover is affected by the related factors and try to build models to predict whether the certain employees will leave the company. As for the practical contribution, this project aims to provide the Human Resource Department guidelines on how to retain valuable employees and prevent a high employee turnover. Overall, the final goal of this project is to allow the company to take better decision-making actions on human resource management.

In order to analyze the problems in practical details, we apply the case of a mid-sized IT company with over 10,000 employees which is considering how to design a reasonable plan to maintain the valuable employees.

## 2 DATA

### 2.1 Data Information

The data was found from the "Human Resources Analytics" dataset provided by Kaggle's website https://www.kaggle.com/ludobenistant/hr-analytics. There are totally 15,000 training examples in this dataset. Each example includes ten variables:

Independent Variables:

- `Employee satisfaction level`
  It is a numerical variable with range between 0 and 1. The larger the value is, the more satisfied the employees are with the company.
- `Number of project`
  It is a numerical variable but also is a count variable limited to a small set of integers. It shows how many projects the employee has already involved in this company.
- `Last evaluation score`
  It is a numerical variable with range between 0 and 1. It shows the performance score of an employee given in his or her last performance review.
- `Average monthly hours`
  It is a numerical variable. It shows the mean working hours of the employee in a month.
- `Time spent at the company`
  It is a numerical variable but also is a count variable limited to a small set of integers. It shows how many years the employee has already stayed in this company.
- `Whether they have had a work accident`
  It is a binary variable. 1 means the employee has suffered work accident in this company and 0 means the employee has not.
- `Whether they promoted in the last 5 years`
  It is a binary variable.1 means the employee has a promotion in the last 5 years and 0 means the employee has not.
- `Department`
  It is a category variable with 10 categories, including sales, accounting, hr, technical, support, management, IT, product, marketing and RandD.
- `Salary`
  It is a category variable with 3 levels, namely low, medium and high.

Dependent Variables:

- `Left`
  It is a binary variable. 0 means the person is still employed by this company (about 77 %). And 1 means the person has already left the company (about 23 %).

### 2.2 Data Preprocessing

We read the data from the csv file and renamed columns. The dataset has 15000 data points and 10 variables. There are two potential problems that might occur in our dataset. One is missing values, and the other is outliers.

*a. Missing Values.*

Missing values occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. We checked if there are missing values in the dataset, and we found that there are no missing values in the dataset.

*b. Outlier Detection.*

Outliers can lead to wrong estimations and poor performance on some of the algorithms. An outlier is a data point that seems to be out of the normal when it comes to a variable or even multiple variables at once. Usually outliers can be univariate or multivariate which means that their abnormal values can either happen on one dimension or multiple variable dimensions together.

In our project, we focused on univariate outliers. The initial value to detect outliers used in this analysis is a 1.5 multiple of the Inner Quartile Range. We found that there are 1,282 outliers in only one variable (time spend company). The outlier is 8.55% of the overall dataset. We then used a broader range to confirm whether the outliers should be removed. After widening the step range by 0.5, the outlier-rate is reduced to less than 3.5 %. Therefore we decided that all possible outliers of this variable will be used for the analysis.

### 2.3 Data Explorations

*a. Exploration on dependent variables.*

The variable left is dependent variable and it's a binary variable. We calculated the turnover rate (the rate of staff who left among all staff), and it's about 0.24. This shows that the data samples are imbalanced, which may cause problems in the classification prediction. The details of addressing the imbalanced data problem will be given in the following sections.

*b. Exploration on independent variables.*

Among the independent variables, department and salary are categorical variables, work_accident and promotion_last_5years are binary variables, and the others are numerical variables.

For the binary variables, we calculated the rate work accident rate and the promotion rate. The rate of staff who had work accident among all staff is about 0.14. The rate of staff who had promotion in last 5 years among all staff is about 0.02.

For the numerical variables, we calculated the statistics:

| | satisfaction_level | last_evaluation | number_project | average_monthly_hours | time_spend_company |
|---|---|---|---|---|---|
| count | 14999.000000 | 14999.000000 | 14999.000000 | 14999.000000 | 14999.000000 |
| mean | 0.612834 | 0.716102 | 3.803054 | 201.050337 | 3.498233 |
| std | 0.248631 | 0.171169 | 1.232592 | 49.943099 | 1.460136 |
| min | 0.090000 | 0.360000 | 2.000000 | 96.000000 | 2.000000 |
| 25% | 0.440000 | 0.560000 | 3.000000 | 156.000000 | 3.000000 |
| 50% | 0.640000 | 0.720000 | 4.000000 | 200.000000 | 3.000000 |
| 75% | 0.820000 | 0.870000 | 5.000000 | 245.000000 | 4.000000 |
| max | 1.000000 | 1.000000 | 7.000000 | 310.000000 | 10.000000 |

**Figure 1: Statistics for Numerical Variables**

Then we had an overview to summarize the relationship between the dependent variable and the means of the numerical independent variables:

| left | satisfaction_level | last_evaluation | number_project | average_monthly_hours | time_spend_company |
|---|---|---|---|---|---|
| 0 | 0.666810 | 0.715473 | 3.786664 | 199.060203 | 3.380032 |
| 1 | 0.440098 | 0.718113 | 3.855503 | 207.419210 | 3.876505 |

**Figure 2: Relationship between independent variables and dependent variables**

### 2.4 Data Visualization

In this part, some methods are adopted to visualize the dataset to explore it from an intuitive way. We choose package Matplotlib and Seaborn in Python as our tools to curve those diagrams.

We first use heatmap of the correlation matrix to show the relationship between the numerical variables. In the heatmap, the darker the color, the stronger the correlation. The output is in Figure 3, and we can see from the diagram that there is a strong connection between the average monthly hours and the number of project.

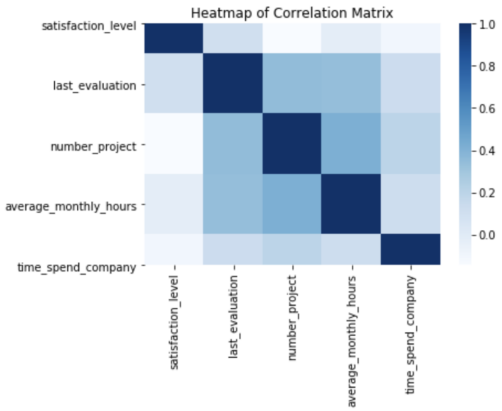


**Figure 3: Heatmap Diagram of Correlation Matrix**

For these three numerical variables: Employee satisfaction level, Last evaluation score and Average monthly

hours, we make the distribution plots of them about the number of employee. The output is in Figure 4.

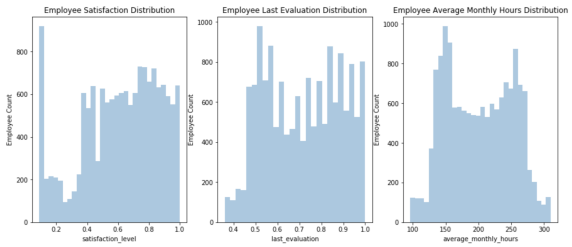


**Figure 4: Numerical Variables Distribution**

We can discover that:

a. There is a huge number of employees with low satisfaction (almost 0.1) and many of the employees are fairly satisfied with the company(more than 0.6).

b. There is a bimodal distribution of employees for low evaluations and high evaluations and also with lower and higher average monthly hours. They have similar shape of diagram.

In order to explore more detail information about these 3 variables, we also draw the kernel density estimation(KDE) diagram for the data points in terms of these 3 variables with label 'turnover' and 'no turnover' respectively. The output is in Figure 5.

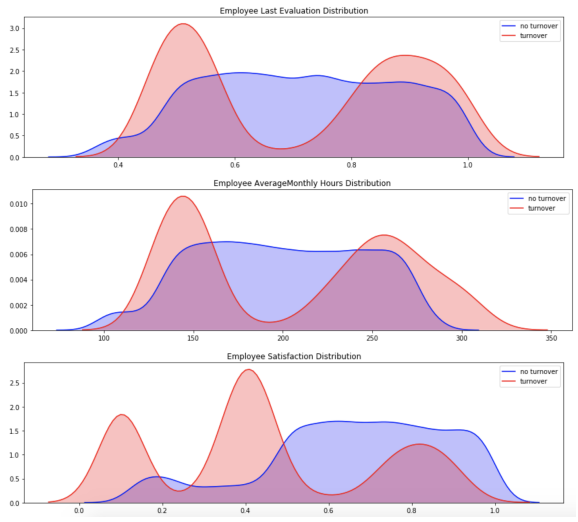


**Figure 5: Numerical Variables KDE Distribution**

We can see that there is a tri-modal distribution for employees who left in terms of satisfaction levels. Employees who had satisfaction levels at around 0.1, 0.4 and 0.8 left the company more. And for the other two variables, the last evaluation and average monthly hour graphs share a similar distribution. From the graphs, we can get an intuition that employees with lower average monthly hours usually got lower score in the evaluation, and they are more likely to leave the company.

## 3 ADDRESSING THE IMBALANCED DATA

Sometimes high accuracy doesn't necessarily mean that the model is adequate. One possible explanation is that the dataset we use is quite imbalanced, or in other words, there exists one class in the training set dominating the other. In this case, the data is not in balance (only 23% of the examples is labeled as leaving the company and the rest is labeled as still staying at the company). If we use the original data to train our models directly, it is possible that the model will tend to predict the vast majority of examples as staying at the company. In the most extreme case, the model can predict all of these examples as staying at the company and has 77% training accuracy, which results in a useless model.

There are multiple methodologies of handling unbalanced data sets in order to deal with this problem. In this project, we mainly focus on modifying class weights and SMOTE (Synthetic Minority Over Sampling Technique) based oversampling techniques.

### 3.1 Class-weight

A useful setting in machine learning models of sklearn is 'class_weight = balanced', wherein classes are automatically weighted inversely proportional to how frequently they appear in the data. i.e. $w_j = \frac{n}{k n_j}$, $w_j$ is the weight to class $j$, $n$ is the number of total observations, $n_j$ is the number of observations in class $j$, and $k$ is the total number of classes.[5]

### 3.2 SMOTE

As an oversampling method, SMOTE synthesizes new minority instances between existing (real) minority instances. It works by selecting two or more similar instances (using a distance measure) and perturbing an instance one attribute at a time by a random amount within the difference to the neighboring instances. We can imagine that SMOTE draws lines between existing minority instances, and synthetic minority instances are just somewhere on these lines. In this way, minority instances could be easily generated without losing important information. [2]

## 4 BUILDING MACHINE LEARNING MODELS

In this section, we apply different machine learning models on our dataset to predict whether the employee will leave the company or not.

### 4.1 Measurements to the models

In Machine Learning, performance measurement is an essential task. The AUC - ROC Curve is one of the most important evaluation metrics for visualizing and checking any classification model's performance. The ROC curve is plotted with TPR (true positive rate) against the FPR (false positive rate) where TPR is on y-axis and FPR is on the x-axis, and AUC represents area under the curve,

telling us how much our classification model is capable of distinguishing between classes.

With imbalanced classes, it's easy to get a high accuracy without actually making useful predictions, therefore in our project we select AUC - ROC Curve as the major measurements to the models instead of using the accuracy.

### 4.2 Machine Learning Methods Used

- Linear SVC
- Logistic Regression
- Decision Tree Classifier
- Random Forest
- AdaBoost

### 4.3 Results and Findings

We build the five machine learning models with three training datasets: the original biased dataset, the original dataset with class-weight and the resampled dataset with SMOTE. The results of each model with different dataset and our findings are listed in the following.

*ROC curve.*

We draw the ROC curve of each model for data with class-weight and data with SMOTE respectively. From the curve, we don't find obvious difference between these two methods. As for models, we can see that Linear SVC and logistic Regression have worst result, while decision tree and random forest model have great performance.

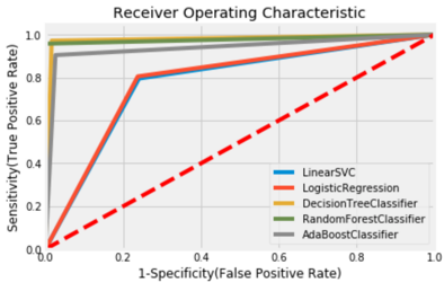


**Figure 6: ROC curve for classification with class-weight**

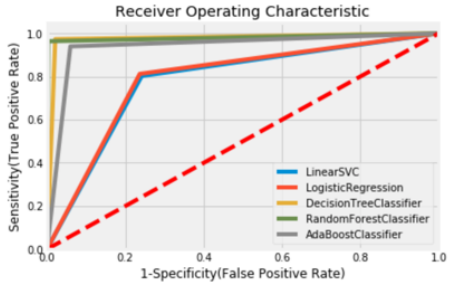


**Figure 7: ROC curve for classification with SMOTE resampling**

*AUC score.*

We compare the AUC score of models with original biased training dataset and with resampled datasets. We find that the linear SVC and logistic regression perform badly (59.92%, 64.23% respectively) on the original dataset, and the resampling methods help to improve a lot (up to around 78% both). The other three models have great performance (higher that 90%) with original dataset, but resampling methods still help to improve the performance. Among the five models, random forest model always has the highest score.

AUC score of each model

| | Original Result | With Class-Weight | With SMOTE |
|---|---|---|---|
| Linear SVC | 59.92% | 77.75% | 78.03% |
| Logistic Regression | 64.23% | 78.39% | 78.78% |
| Decision Tree | 97.02% | 97.79% | 97.62% |
| Random Forest | 97.83% | 97.83% | 98.05% |
| AdaBoost | 93.95% | 93.95% | 94.01% |

*Cross-validation Performance.*

Cross-validation assesses the generalization performance of a machine learning model and can help to avoid overfitting. After comparing the cv performance, we find that the performance of SMOTE is a little better than Class-weight. Among the models, random forest is still the best.

CV performance of each model

| | With Class-Weight | With SMOTE |
|---|---|---|
| Linear SVC | 82.56% | 83.18% |
| Logistic Regression | 82.63% | 83.25% |
| Decision Tree | 97.62% | 97.96% |
| Random Forest | 99.02% | 99.74% |
| AdaBoost | 98.09% | 98.49% |

## 5 BUILDING AN EXPLAINABLE LOGISTIC REGRESSION MODEL

In the section 4, we have already obtained several successful models to predict whether the employee will leave company or not. However, in our project we not only want to know whether the employee will leave, and we also wonder how likely an employee will leave. Besides, we want to know how the possibility of turnover is affected by each feature. In order to achieve the above tasks, we will try to build an explainable logistic regression model. We can obtain the possibility of leaving for given employees using the model and the coefficients of each features in the model can provide us the information about the effect of the features to the leaving probability.

### 5.1 Remove some features

In the above logistic regression models, we use all the features in the dataset to build the model. If there are too many features in the model, it will be difficult for us to interpret the model. In order to reduce the number of variables in the model, we display the Decision Tree Classifier and only select the top five important features to rebuild the logistic regression model. (namely satisfaction_level,

time_spend_company, number_project, last_ evaluation, average_montly_hours)

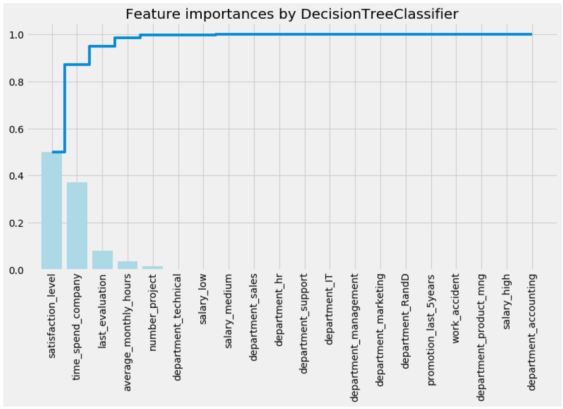


**Figure 8: Features sorted by the importance**

### 5.2 Transformation on the feaures

As showed in the data description part, the value in variables: 'time_spend_company'and 'number_project'is not continuous real number. They are actually a set of limited integer numbers. Such kind of variables can be harmful to the logistic regression model. However it is not suitable to transform the set of limited integer numbers into one-hat vectors directly, because that will bring a quite high-dimensional dataset.

In order to address the problem, having observed these two variables, we have the idea to do the cluster on these two features respectively according to the frequency of the value. In details, we cluster the value in the 'time_spend_company' into 3 classes, namely long, middle and short, and cluster the value in the 'number_project' into 3 classes little, middle and many. Then we split the data into training set and testing data. In order to make the training data balanced, we also conduct SMOTE sampling on the training data.

```
hr_data.head(10)
```

| | satisfaction_level | last_evaluation | number_project | average_montly_hours | time_spend_company | left |
|---|---|---|---|---|---|---|
| 0 | 0.70 | 0.59 | middle | 138 | middle | 0 |
| 1 | 0.72 | 0.70 | middle | 238 | short | 0 |
| 2 | 0.91 | 0.97 | many | 183 | middle | 0 |
| 3 | 0.21 | 0.85 | many | 285 | long | 0 |
| 4 | 0.60 | 0.95 | many | 164 | long | 0 |
| 5 | 0.63 | 0.59 | many | 249 | short | 0 |
| 6 | 0.90 | 0.97 | many | 239 | long | 0 |
| 7 | 0.77 | 0.76 | many | 263 | long | 0 |
| 8 | 0.63 | 0.52 | many | 209 | long | 0 |
| 9 | 0.62 | 0.51 | middle | 222 | middle | 0 |

**Figure 9: First 10 examples of training data after transformation**

### 5.3 Fit the model in R

As we can see, after doing the transformation, 'number_project' and 'time_spend_company' are ordinal variables, so it is not suitable to transform them into one-hat vector and then feed into Sklearn-Logistic Regressor directly. Unfortunately we cannot find a mature package in Python to deal with the regression with ordinal variables. Therefore, we turn to feed the training data into another data analysis-R, and use the function 'glm' in R to build the logistic regression model. While we build the model in R, we transform these two features into factor, then R will treat them as ordinal variables automatically.

### 5.4 The model result

The model result is shown below:

```
Coefficients:
                                    Estimate
(Intercept)                       -6.5090527
number_projectmany                 1.0871766
number_projectmiddle              -1.1710424
average_montly_hours               0.0156352
exp_in_companymiddle               5.3249410
exp_in_companyshort               -1.1840426
last_evaluation                    4.3637829
satisfaction_level                -1.5882934
```

**Figure 10: The coefficients in the model**

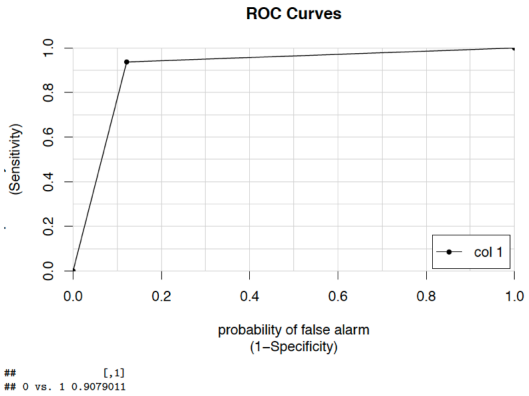


**Figure 11: The AUC-ROC Curve**

The AUC of the new logistic regression is even higher than all above logistic regression models, which convinces us that our transformation on the original data is reasonable and the model is reliable. Then we can use the coefficients in the model to explain how the possibility of turnover is affected by each feature.

### 5.5 Interpretation to the Coefficients

With all of these coefficients, we can know whose employees probably left.

- Employees generally left when they are overworked: $\frac{odds_{manyprojection}}{odds_{littleprojection}} = exp(1.08) \approx 3$.
- Employees with lower satisfaction level was at risk of leaving the company:for 1-unit increase in satisfaction level, odds will multiply by $exp(-1.59) \approx \frac{1}{5}$.

- Employees that had middle experience should be taken into consideration for high turnover rate: $\frac{odds_{middleexp}}{odds_{longexp}} = exp(5.32) \approx 204$ and $\frac{odds_{middleexp}}{odds_{shortexp}} = exp(5.32) \times exp(1.18) \approx 665$.
- Working overtime will make employees has higher odds to left: for 1-unit increase in monthly hour, odds will multiply by $exp(0.015) \approx 1.02$ which is small, but if overtime 2hours/day, the odds of left multiply by $exp[(0.0.15)^{60}] \approx 2.47$.

## 6 DESIGN A MAINTAINING PLAN FOR DIFFERENT EMPLOYEES

Using the logistic regression model we built in section 6, we can predict the possibility of leaving for a given employee.

For example:

| id | satisfaction_ level | last_ evaluation | number_ project | average_ montly_hours | time_spend_ company |
|---|---|---|---|---|---|
| 1 | 0.41 | 0.50 | little | 153 | middle |
| 2 | 0.72 | 0.7 | middle | 238 | short |

$$p_1(leaving) = \frac{1}{1 + e^{-(-6.51-1.59*0.41+4.36*0.5+0.016*153+5.32)}} = 0.94$$

$$p_2(leaving) = \frac{1}{1 + e^{-(-6.51-1.59*0.72+4.36*0.7+0.016*238-1.17-1.18)}} = 0.041$$

This possibility can be used to help the HR department to design retention plan for employees having different possibility of leaving:

- Safe Zone (Possibility score<20%) – Employees within this zone are considered safe.
- Low Risk Zone (20% <Possibility score<60%) – Employees within this zone are to be taken into consideration of potential turnover. This is more of a long-term track.
- Medium Risk Zone (60% <Possibility score<90%) – Employees within this zone are at risk of turnover. Action should be taken and monitored accordingly.
- High Risk Zone (Possibility score>90%) – Employees within this zone are considered to have the highest chance of turnover. Action should be taken immediately.

## 7 CONCLUSION

In this project, we have built some successful models to make prediction. However, besides that we also did many additional jobs to attempt to explain how the turnover is affected by these factors, including using some visualization techniques and taking the advantages of logistic regression model in terms of interpretability. The reason why we do those jobs is that we hope that the output of this project is not only just a theoretic solution using some data analysis techniques, but also it can contribute some new findings and ideas for human resource professionals to solve the employee turnover problem in their organizations.

## 8 ACKNOWLEDGE

In this project, we worked tightly to process the jobs. In details, Zhili Wang is responsible for the data pre-processing and addressing the imbalanced data problem; Song Ma and Zihan Zhang are responsible for the data exploration and data visualization; Xuantong Liu is responsible for building the prediction models; and Weiyan Xie is responsible for building the explainable logistic regression model. We all did the coding job for our responsible part and wrote the corresponding part in the final report.

## REFERENCES

[1] Allen, D. G., K. R. Moffit, and K. P. Weeks. (2005). *Turnover Intentions and Voluntary Turnover: The Moderating Roles of Self-Monitoring, Locus of Control, Proactive Personality, and Risk Aversion*, Journal of Applied Psychology Vol 90 (980990).

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Oversampling Technique.* Journal of Artijcial Intelligence Research, 16:321-357.

[3] Iqbal, Dr. Adnan. (2010). *Employee Turnover: Causes, Consequences and Retention Strategies in Saudi Organizations*, The Business Review, Cambridge. 16. 275-282.

[4] Punnose, R., and A. Pankaj. (2016). *Prediction of Employee Turnover in Organizations Using Machine Learning Algorithms*, International Journal of Advanced Research in Artificial Intelligence Vol 5.

[5] https://chrisalbon.com/machine_learning/logistic_regression/handling_imbalanced_classes_in_logistic_regression/