# **Applying Survival Analysis to study Employee’s Turnover**

By : Houda LABI & Clément COSTE

The goal of this report is to predict an Employee's risk of quitting their job (with a Survival Analysis Model). First, we cleaned and explored the dataset to prepare it for analyses. After reviewing the statistical organization's and performance characteristics through EDA and feature engineering techniques, we generate models that predict employee turnover.

This notebook predict the Turnover Employee's, using the workflow bellow:

1) Data Import and Clean (‘Turnover.csv’)

2) EDA

3) Feature Engineering

4) Significance Testing

5) Logistic Regression Modeling and Validation

**Dataset**

This Employee Turnover dataset is a real dataset shared from Edward Babushkin's blog used to predict an Employee's risk of quitting a job.

Our Dataset (turnover.csv), contains 16 variables, as bellow:

**Column Attributes**

* **stag** - Experience (time)
* **event** - Employee turnover
* **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 pipelene employee came to the company.
* **coach** - Presence of a coach (training) on probation
* **head\_gender** - head (supervisor) gender
* **greywage** - The salary does not seem to the tax authorities. Greywage in Russia or Ukraine means that the employer (company) pay
* **way** - Employee's way of transportation
* **extraversion** - Extraversion score
* **independ** - Independend score
* **selfcontrol** - Selfcontrol score
* **anxiety** - Anxiety score
* **novator** - Novator score

Turnover employee’s is an interesting subject for compagnies. Studying this subject will give us more informations about employee’s quitting their jobs and try to give solutions to compagnies in order to reduce the turnover. However, there’s not much written about this subject in documentation and internet, that’s why we choose this dataset, in order to be more challenged.

In this reposrt we will explain a part of turnover analysis using the survival analysis and using a real dataset.

We studied data by coding in R and Python for a final part for model Linear regression:

**CODE in R:**

Necessary packages:

Text

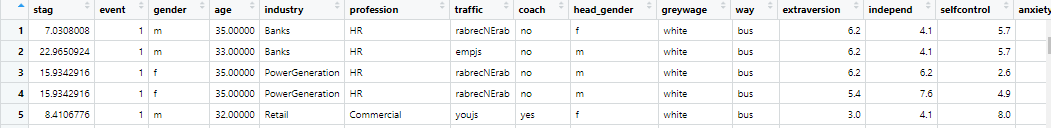
Description automatically generated

Load the dataset (turnover.csv):

A picture containing chart

Description automatically generated

Let’s have a look on data:

…

We get a data dimension of 1129 rows and 16 columns (see R file)

**Variables/Distribution:**

In this dataset, we have Continuous variables, such as:

event = (Turnover or active), stag (time to turnover in months), Age (employee’s ages), Independence(level of employee independence), extraversion (level of employee extraversion), selfcontrol (level of employee selfcontrol), Anxiety ((level of employee anxiety), Novator (level of employee novation).

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

And Discret variables, such as:

Gender (Male or Female), Manager gender(f or m), Coach(No coach, Yes or my head) , greywage (White or grey), traffic(which traffic employee takes), Profession, Transport(car, bus or foot) and Industry.

Chart, bar chart

Description automatically generatedChart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, shape

Description automatically generated

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

Chart

Description automatically generatedShape, square

Description automatically generated

Note : see link for salary greywage : https://blogs.elenasmodels.com/ukraine-white-black-grey-wages/

Grey wage = Employees are paid cash partially, in addition to the official wage (“grey” wages”).

White Wage: employees receive “white” (fully official) wages.

See code in R Script.

#Explanatory of Employee's Turnover VS active employee's:

Chart, bar chart, treemap chart

Description automatically generated

The graph above shows status employee's (Turnover or still active).

The graph indicates that there are 558 active employees and 571 inactive employees (Turnover done) in the dataset (n = 1129).

**Correlation table:**

stag event age anxiety extraversion novator selfcontrol independ

stag 1.0000000000 -0.048360700 -0.19738094 0.01475463 -0.08822668 -0.037632684 0.07707563 0.0005504976

event -0.0483606998 1.000000000 -0.04875073 -0.06323155 0.01545827 0.006825492 -0.04003960 0.0518639135

age -0.1973809390 -0.048750730 1.00000000 0.05778191 -0.14975284 0.039509274 0.03899615 0.0561289192

anxiety 0.0147546292 -0.063231548 0.05778191 1.00000000 -0.13504591 0.246667673 -0.10756847 -0.4272090041

extraversion -0.0882266752 0.015458270 -0.14975284 -0.13504591 1.00000000 0.297374814 -0.53803886 -0.2000521587

novator -0.0376326844 0.006825492 0.03950927 0.24666767 0.29737481 1.000000000 -0.56597244 0.0238651461

selfcontrol 0.0770756294 -0.040039603 0.03899615 -0.10756847 -0.53803886 -0.565972436 1.00000000 -0.1657952876

independ 0.0005504976 0.051863914 0.05612892 -0.42720900 -0.20005216 0.023865146 -0.16579529 1.0000000000

We do not see a good correlation between those features.

**Distribution on time:**

Chart, line chart

Description automatically generated

Looking at the graph ‘ Distribution on time’ we can tell that experience (stag = time) has a very weak correlation with employee quitting so we can tell that experience (time) is not a major factor on employee quitting.

**GENERAL Kaplan Maier without covariates:**

Kaplan-Meier test is used in our case to calculate the probability to stay in a company with and without features:

Chart

Description automatically generated

**Kaplan Maier with Covariates:**

1**- AGE**: we decide to create 4 groups less than 30 years for Young employee, less than 40 for middle age, Less than 50 for Old Age and more than 50 years.

Chart, histogram

Description automatically generated

We can conclude that there is a significant difference between one of the 4 groups --> Pvalue=0.02 <5%

But we can see that there is an interesting difference in the middle age -> Let's reduce groups to two groups (Less than 35 and more than 35Y)

We can also see that the more than 50 years old (have no enough data)

We can also see that the more than 50 years old (have no enough data)

2nd Group, contain a group aged less than and equal to 35Y and another group of more than 35.

Chart, line chart, histogram

Description automatically generated

The pvalue = 1e-07 < 5%, there is a significant difference between the 2 ages groups (Less and more than 35Y)

we going to keep those 2 categories ages to study the rest of data.

**2- No significant parameters:**

We apply a KM on all covariates, and had the results bellow: (Code in R script)

* **Gender :**

p value = 0.1 >5% the difference between gender group is not significant = The sex of employee does not change the probability of turnover

* **Manager gender:**

P value =0.3 >5% No significant parameter.

Same us gender and manager gender**,** **Novator**, **greywage** and **traffic**.variables are not significants. Let’s not consider them in the rest of study.

**3- Profession:**

Chart

Description automatically generated

p= 0.009 is sifnificant -> But not easy to study 16 professions

Let's try to analyse professions and find a relations

According to the general KM profession model + the previous distribution, we tried to select a group of 6 interesting professions, then study them until have a good result.

After, trying to create several groups of professions, we conclude that IT , others (All professions) and HR have a significant impact on our model (See details on R script).

Chart, histogram

Description automatically generated

p= 0.005 is sifnificant ->

Let's keep this group for next

According to the general KM profession model + the previous distribution, I tried to select a group of 6 interesting profession

**4- Employee’s attitudes:** (**Extraversion, Selfcontrol, Anxiety** and **Independence**):

Like we saw before, the novator columns have no impact on our turnover! But what about the rest of attitudes ?

We use KM to verify the impacts of those columns on turnover, and we get the result bellow:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Extraversion | Independence | Anxiety | Selfcontrol |
| Pvalue | **5e-04** | **0.002** | **0.006** | **0.02** |

We can conclude that the employee’s attitude have an impact on their quitting from their job or not.

**5-Transport:**

Chart, histogram

Description automatically generated

the Transport group (Bus, car and foot) have a

p value of 0.003 so it's significant

According to the general KM profession model + the previous distribution, I tried to select a group of 6 interesting profession

**6-Industry:**

Chart

Description automatically generated

We can conclude that the industry have a big impact on the employee turnover Pvalue =2e-07 <<5%

But need to regroup some industry to have more visible and interpretable conclusion.

According to our previous distribution, we going to create an industry group containing the 6 highest industries with turnover-> Banks, Manufacture,etc, Consult, IT and others.

Chart

Description automatically generated

The pvalue = 3e-06, this industry group have a significant impact on employee's turnover.

There is at least one industry having impact on this group! Let's find it. But need to regroup some industry to have more visible and interpretable conclusion.

We cheked the extrem industries in the plot above, such as the Banks + consult, then etc, manufacture and others.

The Pvalue for Banks VS Consult = 0.6 and Manufacture VS etc VS Others =0.1.

We can conclude that we can group:

* Bank and Consult together = BCA
* Manufacture, ETC and Others = MOE
* Finally, we keep the IT.

That’s give us, the following plot:

Chart, histogram

Description automatically generated

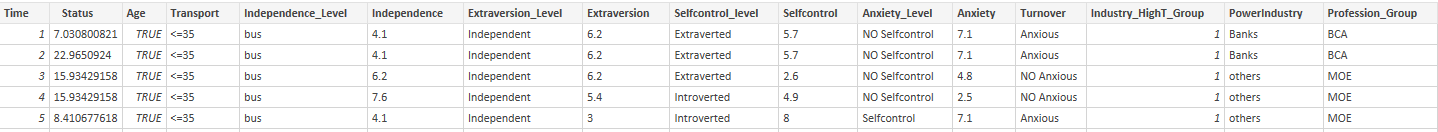
The p= 5e-07 is highly significant ->We have a significant industry groups in our model

we can conclude that our group is successful.

NOTE: After studying our data, we clean it in order to:

* Remove the uninteresting variables,
* Categorize age (less than 35 years and more)
* Categorize attitudes by levels
* Create Professions and Industries groups.

(See code in R script)



Coxph: Proportional Hazards Model:

In this part we use the coxph to confirm our KM models. Let’s check:

**1-Coxph~Age:**

A picture containing text

Description automatically generatedText, letter

Description automatically generated

The cox model with Age covariate is statistically significant for the 3 alternative tests with a pvalue <<<5%

For Age more than 35Y the coef(beta) =0.505, means that employee’s having more than 35 years, have more risk to stay in the same job than employees having less than 35years.

The hazard ration ex(beta) =1.65, means employees having more than 35years increase the hazard by a factor of 1,65 to stay in their job.

Plus, The Pvalue of Age covariate = 1.47e-07, that’s mean that Age have a significant impact on the turnover.

**2-Coxph~ Age + PowerIndustry:**

The 3 alternatives tests gives a significants pvalue p=5e-12 <<<5% = The model is significant

The variable Age (>35) have a highly statistical significant coefficient 3.06e-08 with a hazard ration of 1.71 Indicating a strong relationship between the Age (>35) and increasing the risk of staying in the same job.

In the other way, the pvalue of Industries --> IT or MOE are also <<< than 5% with a HR of 0.4 and 0.6 with a big Confidence interval. = these results indicate that profession (IT and MOE) makes contribution to the difference in the Hazard Ratio after adjusting for the profession values and employees age

Note: In this section, we studied several features combination (See R script).

**REMARK:**

According to our study, we can see that we have a high risk of turnover with employee's having less than 35years ,that are anxious, extraverted, working such an IT, etc...

There exist a lot of parameters impacting the turnover. That’s why an oriented project (ex: Turnover VS salary or Turnover by studying employee’s attitude) is more efficient and going to limit the covariates to study.

For our general view, we decide to create a column of a group of high risk and another one with low risk, with 2 variables (Age and Anxiety):

Note: different combination can be done! but let's focus in chosen one:

* **High risk Group and Low risk group (Age + Extraversion covariates):**

Background pattern

Description automatically generatedChart, line chart

Description automatically generated

the p= 7e-05 of our Risk group is significant! we can conclude that Age and Anxiety are an important factors in turnover employee's.

**Logistic Regression Model (with Python):**

In this section, we export our R ‘Clean\_data ‘to a csv file and read it in Python, in order to create a small model of Logistic regression.

Same us, our R code we going to create a model to predict the turnover (0 or 1) using Age and extraversion covariates.

To do that, we need to encode our data (OneHotEncoding) and create a pipeline between our data, transformed columns and logistic regression model.

Graphical user interface, table

Description automatically generated

(See Python code)

Conclusion: For this kind of studies, we need to have more data and robust features, such as: Salary, the date of hiring, date of turnover, Kind of degrees etc…

Employee’s attitudes, Transport and age are not strong features to predict a turnover.

In the Logistic model part, we tried several combinations of covariates and was not able to have more than 0.55 accuracy.

“Survival analysis” is one of the most important algorithm but it’s not the most popular one to predict employee turnover.

For our chosen data, it’s better to develop more robust logistic regression model, rather than Survival analysis.