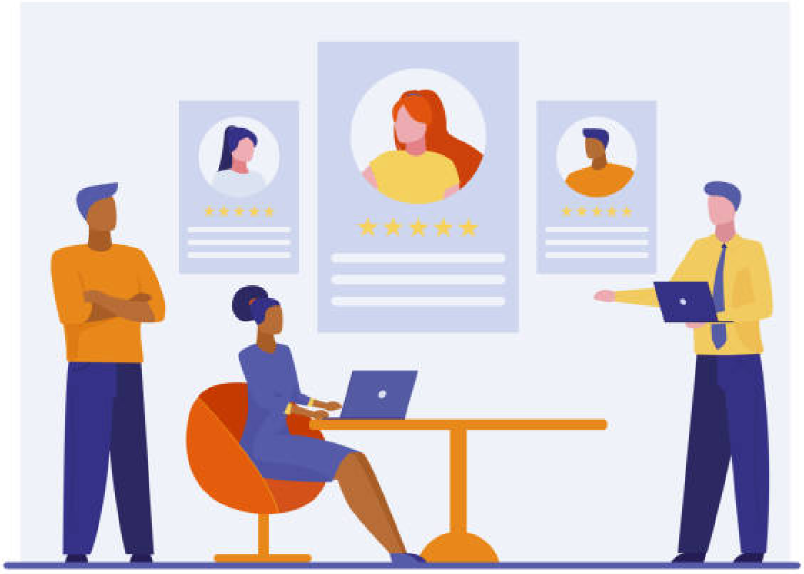


Human Resources data analysis

Predictions about employee churn || Employee survival analysis

Rafayel Mnatsakanyan
Pertchuhi Proshyan
Sona Khloyan

13/12/2022

American University of Armenia

MGMT325 - Business Analytics

Introduction

Short summary

Employee turnover directs to the number of employees who quit a company over a certain period. It contains those who exit voluntarily and employees who are terminated, that is, involuntary turnover.

Voluntary turnover is when employees themselves choose to quit. This can result from more satisfactory job options elsewhere, disagreement within the workplace, disengagement, etc.

Involuntary turnover is when an employer decides to quit an employee, conceivably because of unsatisfactory performance, toxic behaviour, or other reasons.

There are many explanations employees exit a department or an entity, and while some factors for turnover are negative, some turnover is foreseen and completely normal. What's worse is when turnover happens for negative reasons. Some of the most typical reasons for turnover include the following:

- Absence of opportunity for growth or career development
- Career Advancement
- Internal promotion
- Feeling overworked/burnout
- Opposing feelings towards supervisor or leadership
- Unhealthy work environment

Contributions

In the scope of our group project we have chosen an HR Analytics dataset from Kaggle.com uploaded by Giri Pujar in 2018. Based on the observations from the dataset, our main goal will be to predict whether the employee will churn a company or not, which variables will affect it the most and run a survival analysis. This analysis is very interesting and topical for the business as the company's high turnover rate causes instability and unhealthy conditions. So analysing this dataset will contribute to the companies realizing the factors that may lead to employee churn.

Literature review

For this purpose, we will use Decision trees, Survival analysis, Random forests and Bagging algorithms in our analysis.

A decision tree is a distinctive type of probability tree that allows deciding some process. Trees are a perfect way to deal with complex decisions, which always involve many different factors and usually involve some uncertainty. There are two types of Decision trees: Classification trees are used when the dataset needs to be divided into classes that belong to the response variable. Regression trees are used when the response variable is continuous. (Glen, 2019)

Classification Tree Analysis (CTA) is an analytical approach that takes examples of known classes (training data) and makes a decision tree based on measured characteristics such as reflectance. (Clark Labs, 2018)

Survival analysis is a group of statistical techniques for data analysis for which the output variable of interest is time until an event happens. Goals Of Survival Analysis is to assess the relationship of explanatory variables to survival time. (Dhamodharan, 2022)

Bagging is an algorithm that fits multiple models on various subsets of a training dataset and then merges the predictions from all models. (Brownlee, 2020)

Random forest is an extension of bagging, which also randomly determines subsets of components used in each data sample. Both bagging and random forests have been verified as effective on a broad range of diverse predictive modelling issues. (Brownlee, 2020)

Research question

The primary objective of our research is to predict whether the employee will churn or continue working in the company and which factors most influence the decision to churn the company.

The secondary objective of our research is to analyse the survival rates of employees in ascertain company, grouping them by some categories, and understand employees of which group tends to have the lowest and highest survival times in a company.

Our hypothesis is that the decision to churn the company is formulated by some characteristics. Such characteristics could be the satisfaction level an employee has towards the company based on survey results, whether an employee experiences work accidents in a workplace or not and what is the most intuitive the salary an employee is paid.

Our research below may be useful for Human resources managers and Executive managers to create a healthy and comfortable environment for employees to avoid high rates of employee churns.

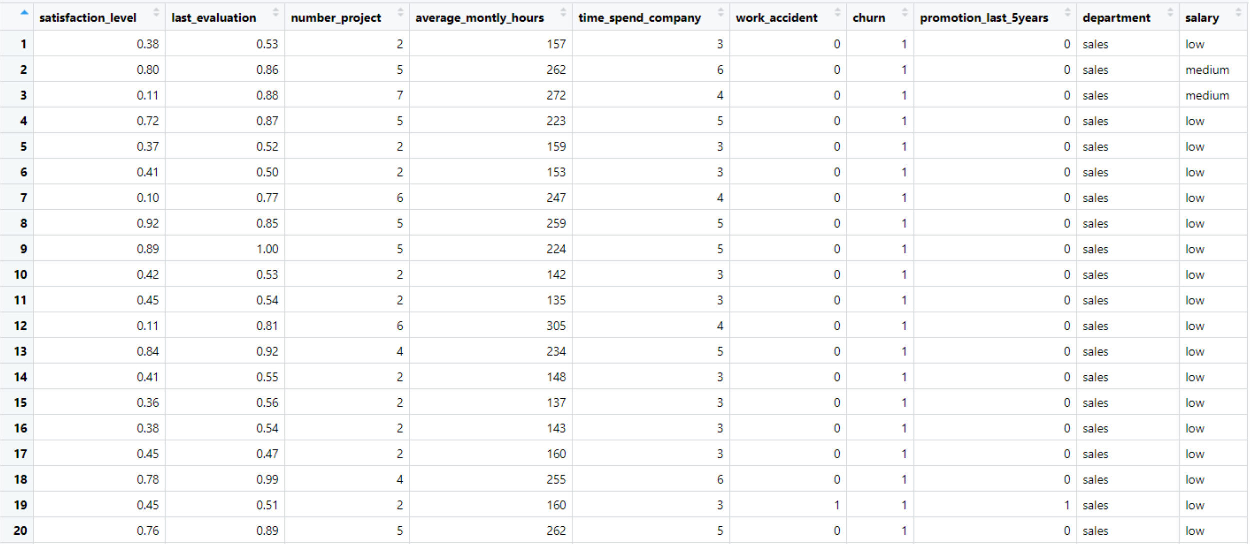
Data

We used the Kaggle.com website to get the dataset for our analysis. We have chosen the HR_Employee_Data.xlsx dataset from the HR Analytics page and later converted the data to CSV format. For our further analysis, we have modified the percentile variables to proportional ones.

This dataset is obtained from the HR department of one of the MNC (Multinational corporation) companies which is not indicated here. The dataset contains eleven descriptive variables and around 15000 rows. The variables are the satisfactory level and last evaluation results of an employee, the number of projects conducted at a time and average monthly hours an employee works, whether the employee had a promotion during the last five years or experienced a work accident or not, whether the employee gets a low, high or medium salary, etc. Find the detailed description of the dataset in the Appendix.

Descriptive statistics (describe the sample)

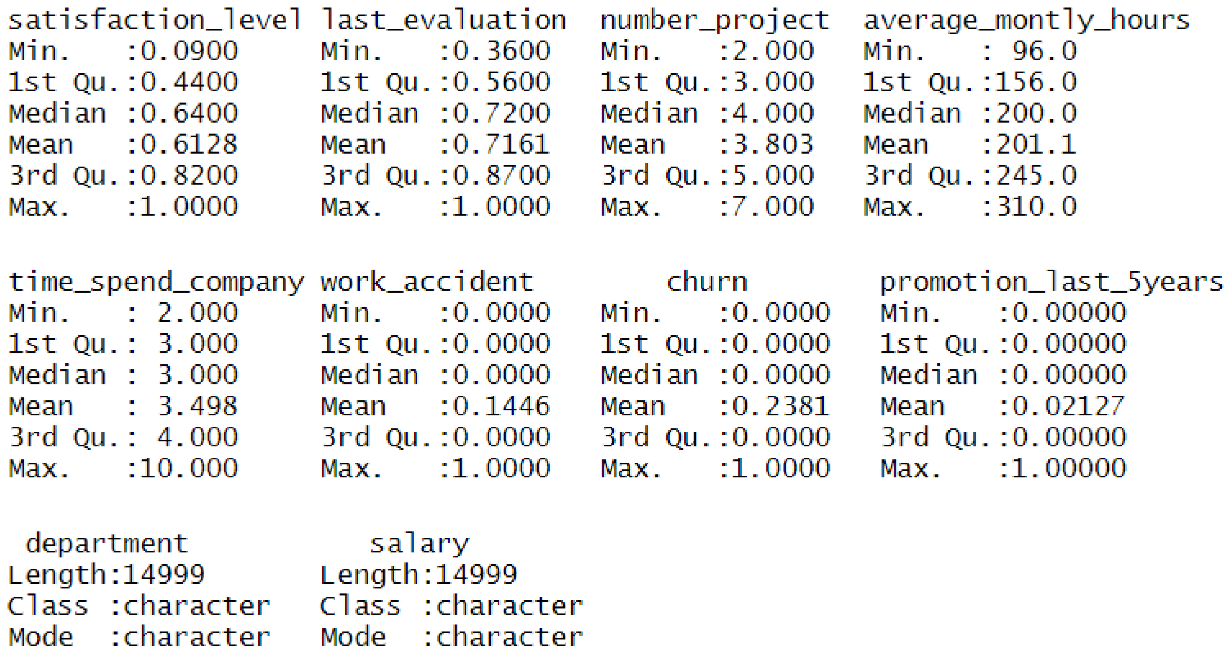
The table below represents the first 20 observations of the dataset about employee turnover on which we've run our analysis:



	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	work_accident	churn	promotion_last_5years	department	salary
1	0.38	0.53	2	157	3	0	1	0	sales	low
2	0.80	0.86	5	262	6	0	1	0	sales	medium
3	0.11	0.88	7	272	4	0	1	0	sales	medium
4	0.72	0.87	5	223	5	0	1	0	sales	low
5	0.37	0.52	2	159	3	0	1	0	sales	low
6	0.41	0.50	2	153	3	0	1	0	sales	low
7	0.10	0.77	6	247	4	0	1	0	sales	low
8	0.92	0.85	5	259	5	0	1	0	sales	low
9	0.89	1.00	5	224	5	0	1	0	sales	low
10	0.42	0.53	2	142	3	0	1	0	sales	low
11	0.45	0.54	2	135	3	0	1	0	sales	low
12	0.11	0.81	6	305	4	0	1	0	sales	low
13	0.84	0.92	4	234	5	0	1	0	sales	low
14	0.41	0.55	2	148	3	0	1	0	sales	low
15	0.36	0.56	2	137	3	0	1	0	sales	low
16	0.38	0.54	2	143	3	0	1	0	sales	low
17	0.45	0.47	2	160	3	0	1	0	sales	low
18	0.78	0.99	4	255	6	0	1	0	sales	low
19	0.45	0.51	2	160	3	1	1	1	sales	low
20	0.76	0.89	5	262	5	0	1	0	sales	low

Table 1

Before knitting to our project for descriptive statistics, we have done some visualisations that describe the main patterns in our dataset.



```
 satisfaction_level last_evaluation  number_project  average_montly_hours
 Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0
 1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0
 Median :0.6400     Median :0.7200   Median :4.000   Median :200.0
 Mean   :0.6128     Mean   :0.7161   Mean   :3.803   Mean   :201.1
 3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000   3rd Qu.:245.0
 Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :310.0

 time_spend_company work_accident        churn         promotion_last_5years
 Min.   : 2.000     Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
 1st Qu.: 3.000     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
 Median : 3.000     Median :0.0000   Median :0.0000   Median :0.00000
 Mean   : 3.498     Mean   :0.1446   Mean   :0.2381   Mean   :0.02127
 3rd Qu.: 4.000     3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
 Max.   :10.000     Max.   :1.0000   Max.   :1.0000   Max.   :1.00000

  department           salary
 Length:14999       Length:14999
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Table 2

The table above is the summary statistics of our data.

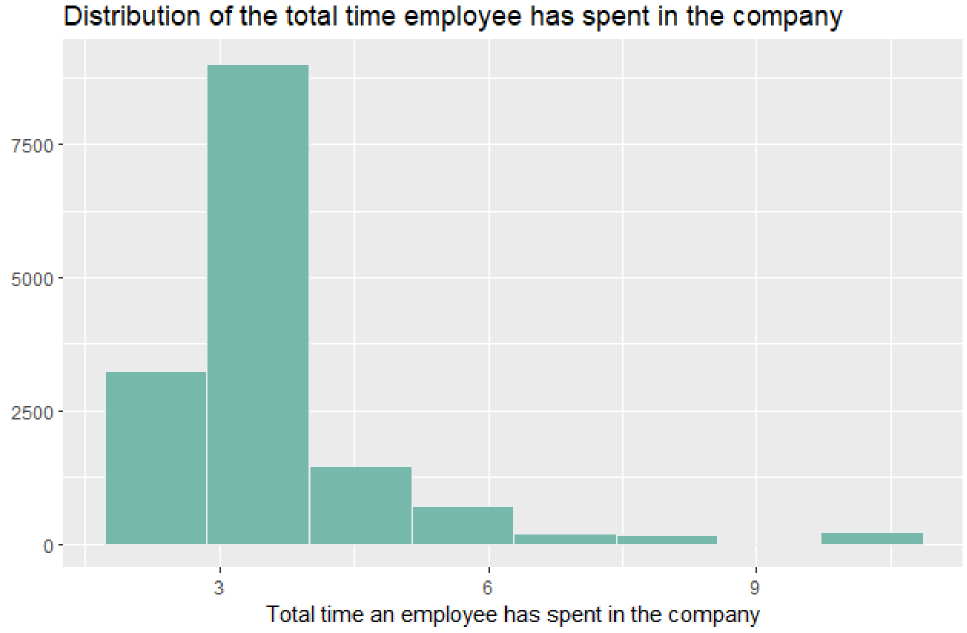


Figure 1

Figure 1 shows and confirms the pattern that employees mostly leave the company after 2-5 years, maybe for personal and career growth. Also, employees do not spend more than ten years in one company.

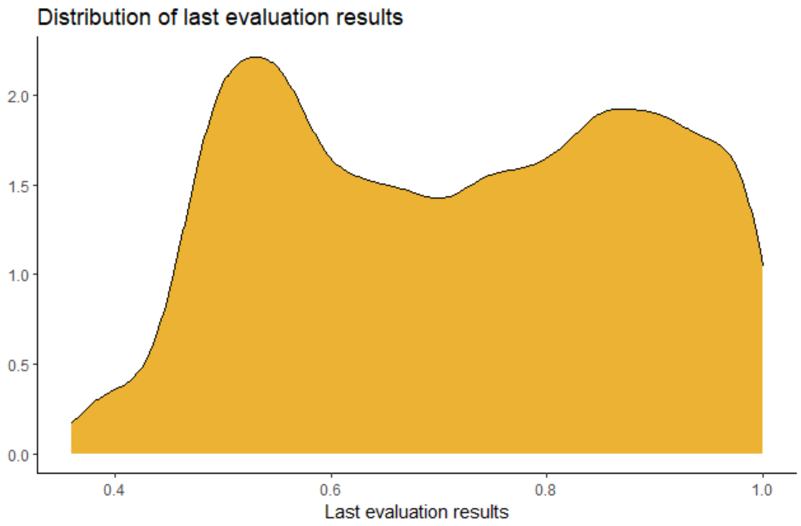


Figure 2

Figure 1 represents the results of the last evaluation results which mostly lay in the range of 0.5 to 0.6. Also, a large number of employees have a result between 0.8 to 0.95.

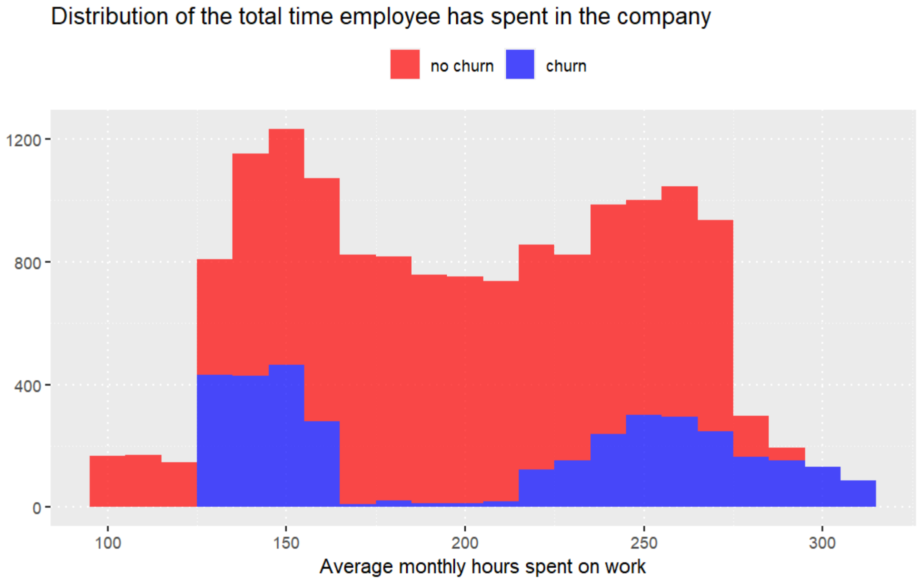


Figure 3

Figure 3 above represents the distribution of average monthly hours employees spend on work. The colouring is done on factor whether the employees churn the company or not. The conclusion is that the more hours employees spend at work, the likelier they'll stay in the company.

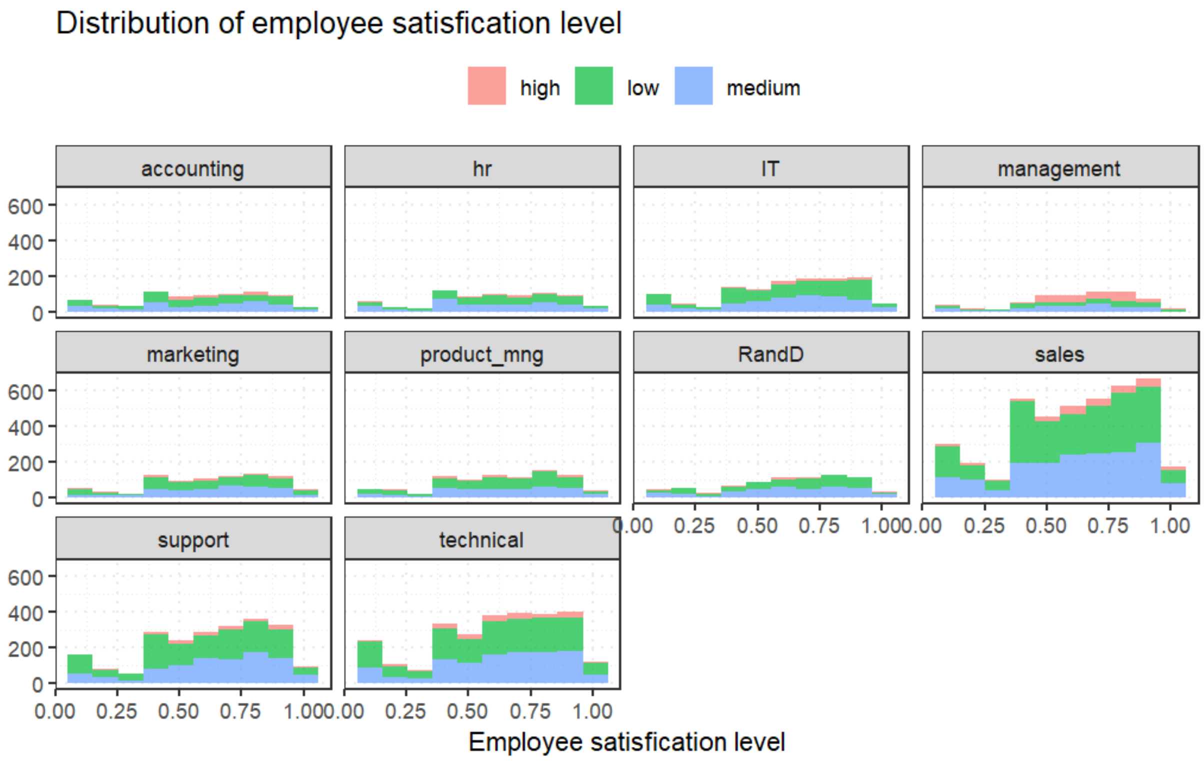


Figure 4

Figure 4 shows the satisfactory level of employees in different departments. Colouring is done based on the salary level whether it is low, medium or high. It is obvious that most people with medium salaries have a low satisfactory level. Also, we see that people in the management department are paid well, high (which is intuitive). Employees in sales, support, and technical departments have a comparatively higher satisfaction level.

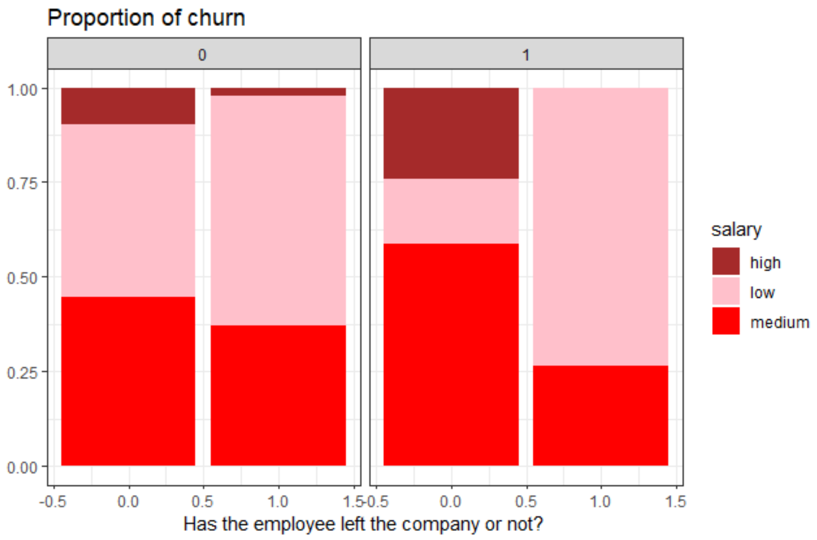


Figure 5

Figure 5 represents the proportion of churns, considering whether the employee had a promotion during the last five years or not. We can see that in both cases, employees with low salaries churn the most. Also, among the ones who do not churn are employees with medium salary rates.

Analysis method

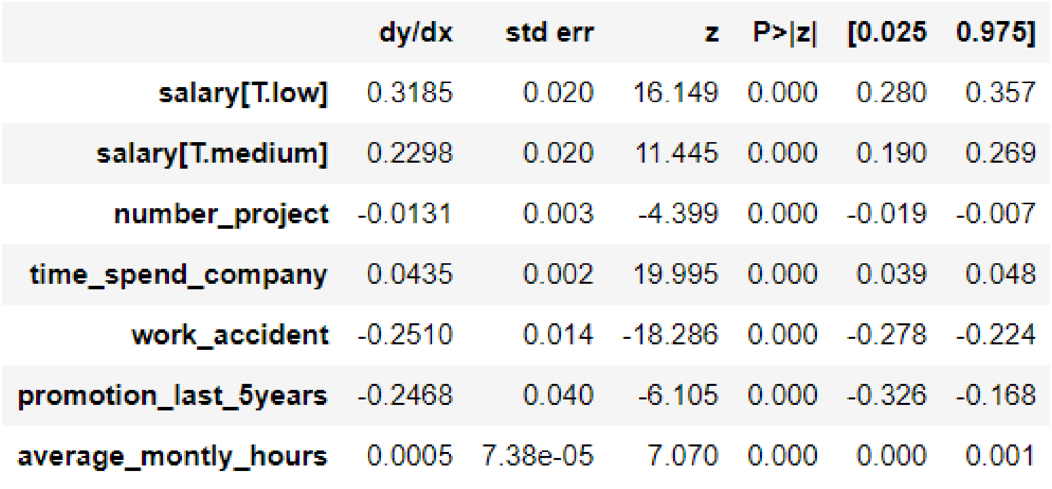
Python 3 has been used to conduct our analysis. The model and libraries which we have run are:

- Pandas
- Matplotlib
- Numpy
- Statsmodels
- Sklearn
- Lifelines

Logistic regression

Initially, our goal is to quantify the factors that affect employee's decision to churn the company. To get an appropriate answer, we have run a logistic regression, which of statistical model and is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring based on a given dataset of independent variables. After, to quantify the effect of changes, we got the marginal effects of independent variable.

Below are represented the marginal effects:



	dy/dx	std err	z	P>|z|	[0.025	0.975]
salary[T.low]	0.3185	0.020	16.149	0.000	0.280	0.357
salary[T.medium]	0.2298	0.020	11.445	0.000	0.190	0.269
number_project	-0.0131	0.003	-4.399	0.000	-0.019	-0.007
time_spend_company	0.0435	0.002	19.995	0.000	0.039	0.048
work_accident	-0.2510	0.014	-18.286	0.000	-0.278	-0.224
promotion_last_5years	-0.2468	0.040	-6.105	0.000	-0.326	-0.168
average_montly_hours	0.0005	7.38e-05	7.070	0.000	0.000	0.001

Table 3

Our conclusion from logistic regression analysis is as follows:

- Unit increase in average monthly hours an employee works is expected to increase the probability of churn by 0.06 %.
- Unit increase in years an employee has spent in a company is expected to increase the probability of churn by 4.13%.
- Unit increase in the number of projects an employee conducts is expected to decrease the probability of churn by 1.2%.
- People getting low or medium salaries have more chance to churn the company than those with high salaries. The probability is increased by 24-33%.
- People having work accidents have 25% less chance to churn the company than those who don't have.
- People who have got promotions during the last five years are 25% less likely to leave than those who haven't got one.

So it is obvious that the amount of salary affects employees’ decision to churn the most: lower-paid employees tend to churn out the company. Also, each additional year spent in the company brings closer employee churn.

Classification tree

Realizing the significant independent variables based on logistic regression, we aim to build a classification tree model. A classification tree is a structural mapping of binary decisions that lead to a conclusion about the class of an object. So, in the case of our analysis, class is about whether the employee will churn the company or stay. We used the satisfaction level and last evaluation results of an employee, the number of projects an employee conducts at a time, how much time the employee has spent in the company and the salary level to see how the churn factor can be predicted based on these values.

Initially, we have split our dataset to test and train parts. The proportion of the test dataset is 10% of the overall dataset. For our first attempt, we have chosen our tree to have two nodes and use gini criterion. Below is the output:

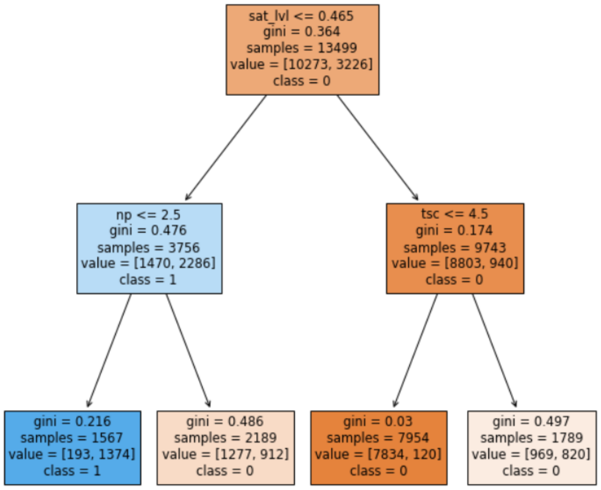


Figure 6

After we calculated the accuracy score of the model using the accuracy_score() function from sklearn.metric module. The accuracy score is the most intuitive performance measure; it is just a ratio of correctly predicted observations to total observations. The higher accuracy score we get, the better our model is. In case of our model, we got around **0.862** accuracy score.

After we calculated the specificity and sensitivity of the model, which are, respectively, the ability of the model to correctly identify people staying in the company and those who will churn. In this model, we got 0.897 for specificity and 0.455 for sensitivity.

Below the feature importances barplot is represented:

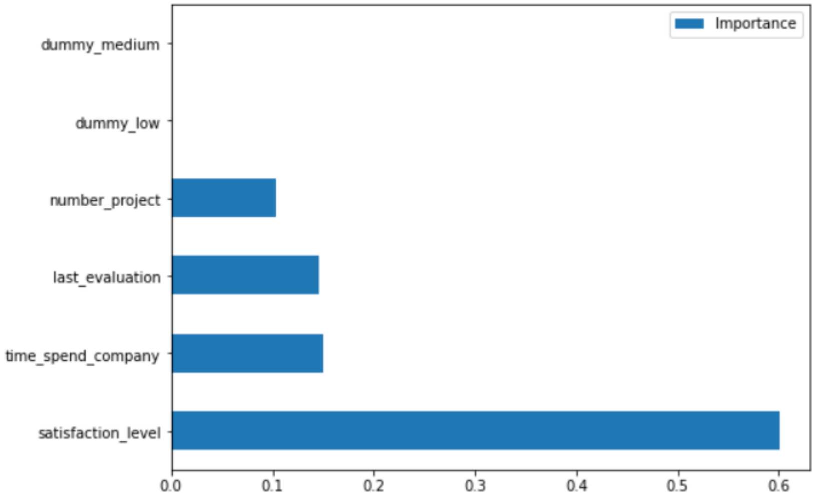


Figure 7

Feature importances plot shows that approximately 60% of the variance of employee churn is explained by satisfaction level, 15% is explained by the time an employee has spent in the company, 14% is explained by the last evaluation results, and 10% is explained by the number of projects an employee conducts at a time.

Results for getting classification tree with maximum depth of 3 you can find in Appendix 2.

Bagging and Random Forest

We use Bagging and Random Forest to improve predictions for the decision tree. These are applied to reduce the variance (overfitting) of the decision tree. Bagging and Random Forest reduce the variance of a single estimate by combining several estimates from different models. As a result, the performance of the model increases, and the predictions are much more accurate and stable.

With Bagging improves our specificity from ~0.897 to ~0.974 and sensitivity from ~0.455 to ~0.971. The feature importances with Bagging are shown in the barplot below.

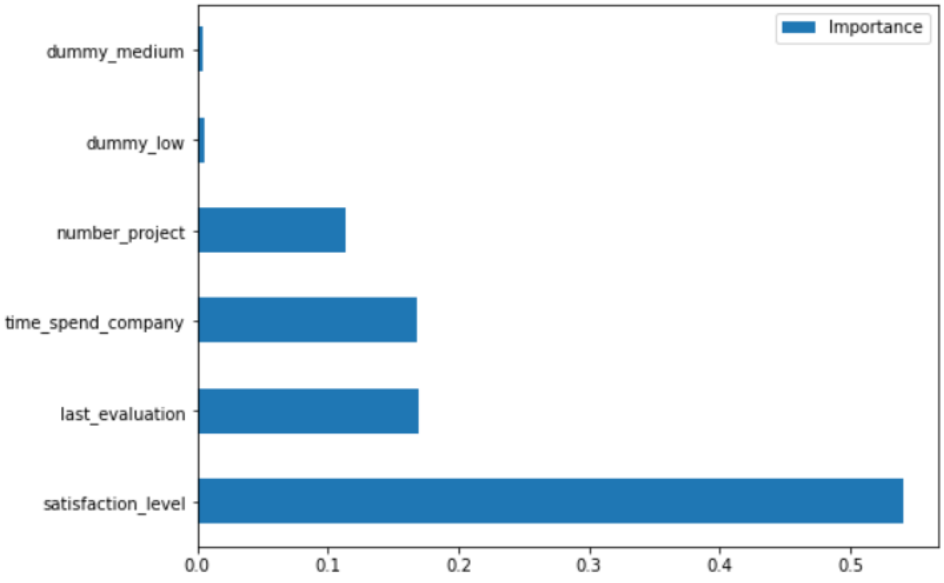


Figure 8

We can see that our dummy_medium and dummy_low variables have gained some importance compared to our initial model.

With Random Forest our specificity and sensitivity change to ~0.977 and ~0.968 accordingly. The feature importances with Random Forest are shown in the barplot below.

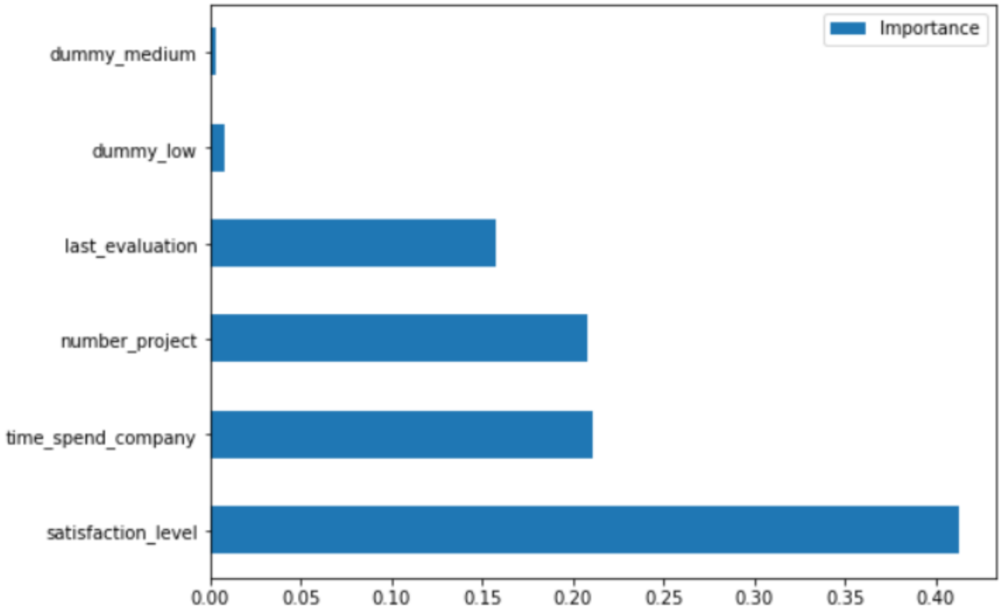


Figure 9

Compared with Bagging in Random Forest dummy_low, time_spent_company and number_project variables' importances increases slightly.

Survival analysis

In the scope of this project, survival modeling deals with the question of how long employees can "survive." Employees are considered to have not survived if they leave the company.

For the survival analysis, we used the Kaplan-Meier estimator method. It is a technique for estimating and plotting the probability of survival as a function depending on time.

In Figure 10, the survival curve is visualized. It is a step function where each drop is caused by an event happening for at least one observation.

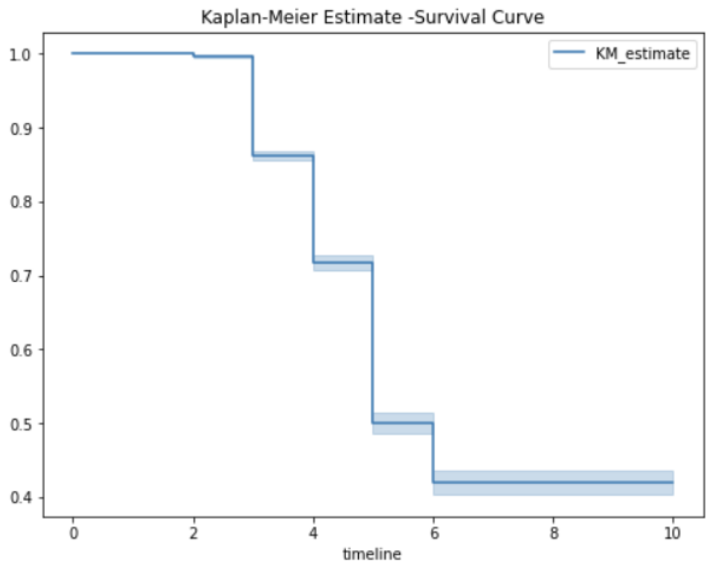


Figure 10

From the graph we can see that the median survival time is 5-6 years which means that on average 50% of the employees have churned after working 5-6 years in the company.

We can also plot the survival curve for different groups in our data. Following graph shows the survival curves for each group of salary.

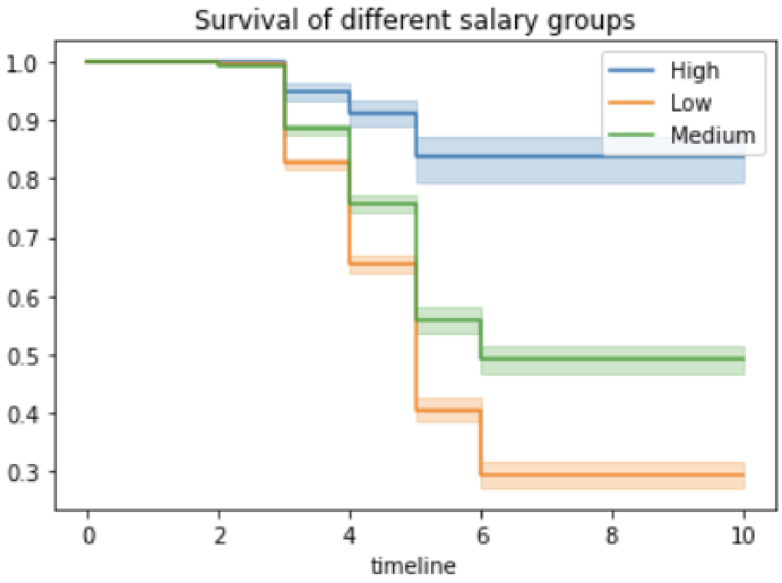
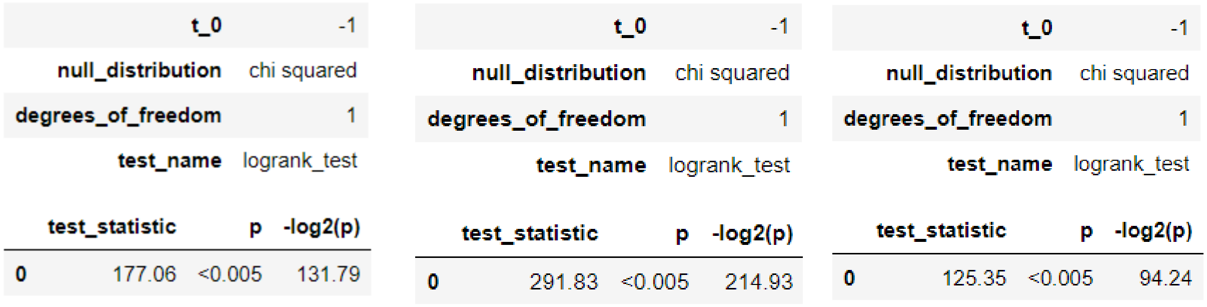


Figure 11

We can see that the probability of survival is the lowest for those who have low salary and is the highest for those whose salary is high.

We can also check their actual difference using the log-rank test. Stating the null hypothesis that there is no difference in survival between 2 groups of interest.



t_0	-1
null_distribution	chi squared
degrees_of_freedom	1
test_name	logrank_test

	test_statistic	p	-log2(p)
0	177.06	<0.005	131.79

Low - medium

t_0	-1
null_distribution	chi squared
degrees_of_freedom	1
test_name	logrank_test

	test_statistic	p	-log2(p)
0	291.83	<0.005	214.93

Low - high

t_0	-1
null_distribution	chi squared
degrees_of_freedom	1
test_name	logrank_test

	test_statistic	p	-log2(p)
0	125.35	<0.005	94.24

Medium - High

Table 4

From the results of log-rank test we can reject the null-hypothesis as the P-value is less than 0.05. So we can state that the survival curves are statistically significantly different.

Results and Conclusion

The result from the analysis we have conducted on HR Analytics data indicates that salary is the most initiative factor to churn the company. People getting low or medium salaries are very likely to churn the company. Besides the salary, the time employee has spent in the company matters as well, meaning that the probability of churn increases yearly. Other factors that affect the employee's churn decision are the number of projects an employee conducts, the average time he/she spends on work, the promotion and work accident factors.

Also, we have conducted survival analysis based on salary groups, and our finding was that the lower salary an employee earns, the lower the survival rate is.

To summarize our analysis, money matters much for any professional individual. Besides the money factor, companies should create a healthy working environment for employees with adequate loads to feel satisfied and appreciated.

References

Glen, S. G. (2019, January 20). *Decision Tree: Definition and Examples*. Statistics How To. https://www.statisticshowto.com/decision-tree-definition-and-examples/

Clark Labs. (2018, June 12). *Classification Tree Analysis*. https://clarklabs.org/classification-tree-analysis/

Dhamodharan, S. (2022, March 30). *Survival Analysis | An Introduction - Analytics Vidhya*. Medium. https://medium.com/analytics-vidhya/survival-analysis-an-introduction-87a94c98061

Brownlee, J. B. (2020, February 12). *Bagging and Random Forest for Imbalanced Classification*. Machine Learning Mastery. https://machinelearningmastery.com/bagging-and-random-forest-for-imbalanced-classification/

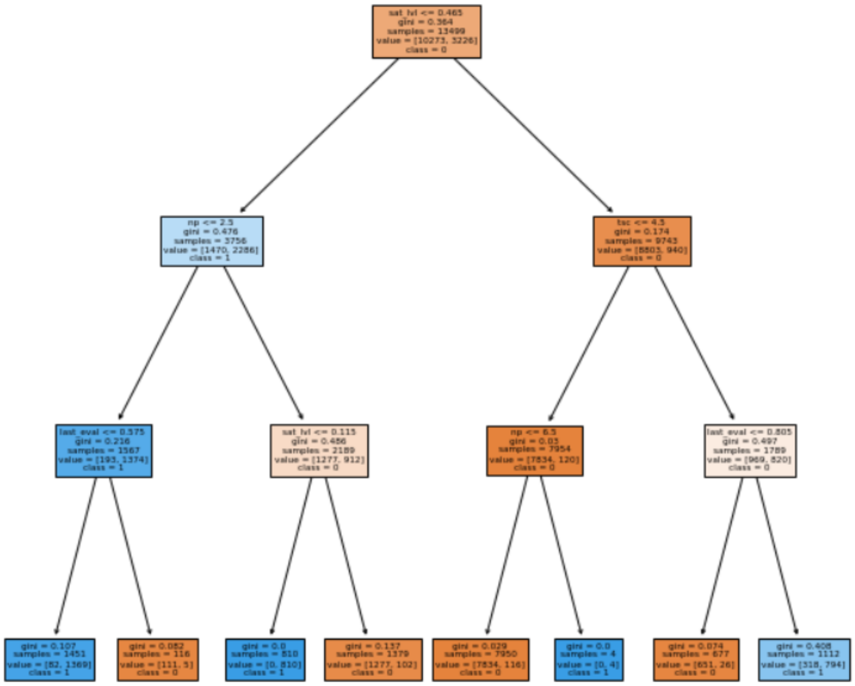
Appendix 1

The table below represents the metadata of our dataset:

Variable name	Explanation
satisfaction_level	Employee satisfactory level.
last_evaluation	Results of the last performance evaluation conducted in the company.
number_project	Number of projects an employee conducts.
average_montly_hours	Average monthly hours an employee spends in the company.
time_spend_company	The total time an employee has spent in the company.
work_accident	Binary variable showing whether the employee has had a work accident in the company or not.
churn	Binary variable showing whether the employee will quit the company or stay.
promotion_last_5years	Binary variable showing whether the employee has got promoted during the last five years or not.
department	In which department the employee works.
salary	Categorical Salary: High, Medium or Low

Appendix 2

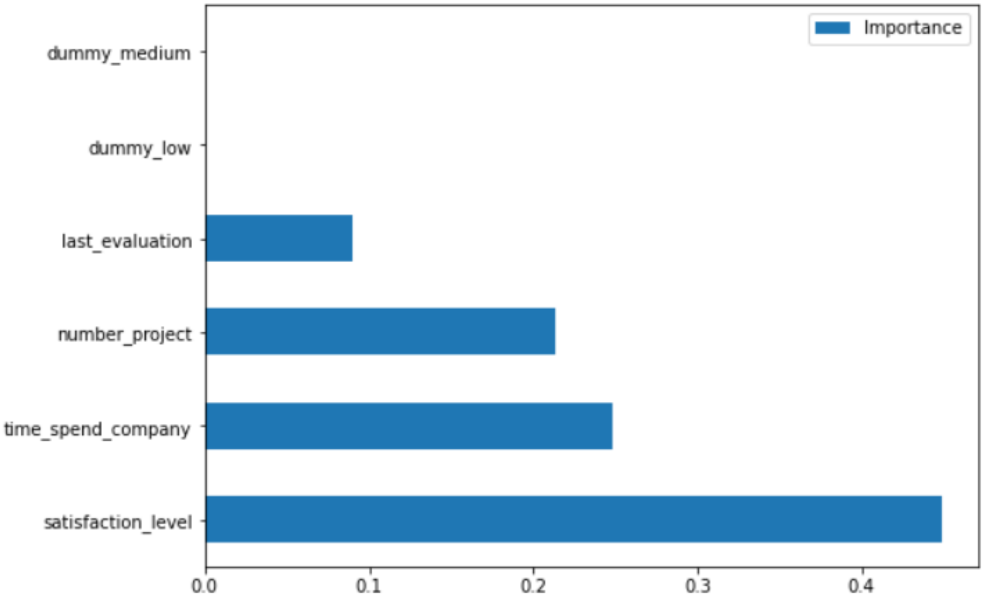
Initially, we have split our dataset to test and train parts. The proportion of the test dataset is 10% of the overall dataset. For our second attempt, we have chosen our tree to have three nodes and used the entropy criterion. Below is the output:



Then, we calculated the accuracy score of this model, which is, in this case, around **0.955.**

After we calculated the specificity and sensitivity of the model, which are, respectively, the ability of the model to correctly identify people staying in the company and those who will churn. In this model, we got **0.892** for specificity and **0.915** for sensitivity.

Below the feature importances barplot is represented:



Feature importances plot shows that approximately 48 % of the variance of employee churn is explained by satisfaction level, 25% is explained by the time an employee has spent in the company, 20% is explained by the number of projects an employee conducts and 10% is explained by the last evaluation results.