# HR for Machine Learning

Elpida-Myrto Koutsoni
*Dept. of Electrical Engineering and Computer Science*
*University of Thessaly (UTH).*
Volos, Greece
elkoutsoni@uth.gr

***Abstract*—Human Resources Management (HRM) is a crucial part of every business playing a vital role in its expentidures, earnings and fame. In recent years, Artificial Intelligence (AI) provides the opportunity to insert intelligence in the HRM by gaining valuable insights from large amounts of data. In this work, we make an overview of the HRM Analytics literature and then work on a real data set of a large company. Moreover, sophisticated machine learning techniques are evaluated on their ability to forecast the probability of an employee to leave or stay in the company. The results indicate that ensemble tecniques, such as XGBoost, LightGBM, CatBoost and Random Forest have excellent predictive and training performance being an excellent choice for HR analytics issues.**

***Index Terms*—Human Resources, Analytics, Machine Learning, Attrition, Turnover, Productivity, Expenses, Recruiting, Training**

## I. Introduction

The management of human resources (HR) falls under the purview of the HR department. An HR manager's primary responsibility is to oversee various aspects of employment, including compliance with labor laws, the hiring process, the management of employee benefits, the organization of employee files and documents, and some facets of recruitment. They act as a liaison between a company's management and its workers. Some of the activities that an HR manager performs are:

- Recruiting, training, counseling, and coaching the staff.
- Preserve a pleasant on-boarding process in order to increase productivity.
- Detect the needs of the employees.
- Supervise and evaluate the work.
- Deal with conflicts such as discrimination.
- Compliance with the regulations and laws.
- Manage personnel relations.
- Utilize innovative methods to design work processes.

Artificial intelligence (AI) is a technique for creating computer programs that mimic human thought, judgment, problem-solving, and perception [1]. AI is a rapidly expanding, versatile technology and appears to have significant future potential. Machine Learning (ML) is a sub-field of AI that utilizes data to solve problems. It is also significant because it aids in the development of new goods and provides businesses with a picture of trends in consumer behavior and operational business patterns. Some of the ML's applications are:

- Maps (traffic alerts)
- Virtual Personal Assistants
- Self Driving Cars
- Fraud Detection
- Translate
- Telecommunication networks optimization (5G/6G)

There are four main training models depending on the kind of existing data:

- Supervised learning – input and output with labels
- Unsupervised learning – the algorithm locates classes and patterns in data in the absence of human interference
- Transfer learning – learn and transfer insights from one model to another
- Reinforcement learning – define the goal and reward when successful/punish when failure

Now that we have comprehended the significance of both HR and ML, it is time to see how the combination of these two dissimilar "worlds" can boost the business revenues by decreasing Capital expenditures (CapEX) and Operational expenditures (OpEX). The union of these two fields is called Human Resources Analytics [2] [3]. Organizations can benefit from HR's usage of Machine Learning in a variety of ways. When businesses focus not only on how to drive revenue but also on how they personnel outcomes like growth, performance, and turnover are influenced, things start to improve. Attrition is not only a time-consuming process, but also a financial burden. Knowing a worker's propensity to switch employers can be useful to the company. Decisions to leave a firm frequently result from thorough planning rather than emerging instantaneously. Utilizing both ML techniques and these information, managers will now be able to prevent situations that may cause short or long term expenses for the company.
The main contributions of this work are the following:

- Literature review of the topic
- Experimenting and evaluating various ML algorithms on real data

The target metric is the "left" variable which depicts whether the employee stayed or left from the company. For the ML analysis we use Python. Specifically, for the models implementation we utilize the keras-tensorflow API and numpy and pandas for the statistical analysis. The rest of the paper is organized as follows. Section II where related work is analyzed. Section III provides information about the system's architecture. Section IV where the evaluation of the models

take place and finally Section V where we discuss about the results and how this work can be expanded.

## II. Related work

After extensive literature overview we demonstrate four significant works.

### A. Introducing HR Analytics with Machine Learning

In this book written by Christopher M.Rosett and Austin Hagerty is vividly depicted, with many real-life instances, how ML has ”invaded” in the HR domain and enhanced the process of managing employees and reaching business goals. In the first chapters the authors introduce basic concepts about HR strategies and Analytics. To adduce an example, *HR Analytics Ikigai* is a Japanese philosophy that aims to answer the question ”What are the requirements for good HR Analytics?”. According to the authors a special blend of knowledge from these four key domains is needed to make decisions based on behavioral data:

- Computing
- Human Behaviour
- Statistics and Research Methods
- Business astuteness

These four specialties, each having a significant value, are combined in order to produce equilibrium in HR Analytics. Fig. 1 illustrates an HR Analytics Ikigai, depicted in a four-way Venn diagram.

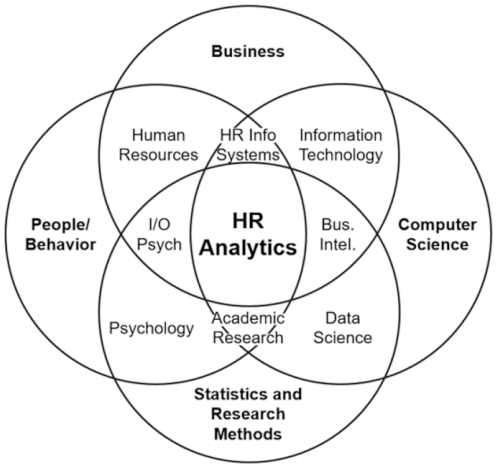



Fig. 1: HR Analytics Ikigai

Another noteworthy point of the book is the analysis of two fundamental reasoning concepts the ”Deductive” and the ”Inductive” [4].

Initially, the *Deductive reasoning* is a kind of logic which through four key actions aims to develop a theory. The four key actions are the following:

- Make observations
- Make hypotheses about the observations
- Scrutinize the hypotheses through predictions and evaluations
- Repeat numerous times in many different but associated experiments

Scientists also know that the theories’ contents are valid because deductive reasoning is used to develop them. They began with a large number of observations and then narrowed the range of potential causes to include the fewest assumptions feasible, leaving them with as nearly unbiased facts as they were able to produce.

Machine Learning introduces now, a new reasoning way, the *Inductive reasoning*. An inductive approach is a bottom-up strategy. It begins with the particular data or patterns and generalises from there to draw conclusions that fit what is visible. The proverb ”construct the bridge while you cross it” applies well to inductive reasoning. A researcher can start down a road using inductive reasoning, then course-correct as they progress.

The main difference between these two concepts is that Inductive reasoning attempts to develop a theory, whereas Deductive reasoning aims to test an existing theory. In real-life business problems timelines are constrict. It is of high importance that decisions are made expeditiously and that is why Deductive method is not always a suitable solution. Using the data at our disposal to deduce patterns and then testing them out is frequently the most plausible strategy when dealing with extremely complicated or little understood systems like the economy or the human mind. Inductive reasoning, as it is depicted in Fig. 2, provides not only pattern recognition but also faster processing and both of them offer the company a competitive advantage.

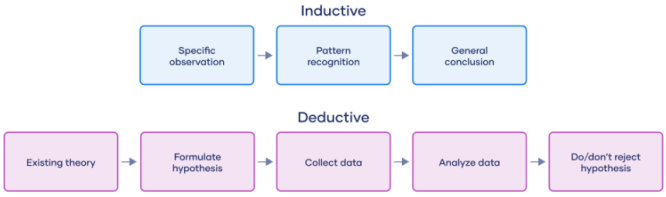



Fig. 2: Deduction vs. Induction reasoning

Subsequently, a detailed introduction of all the fundamental ML methods is taking place. In specific, supervised learning algorithms are presented, such as Linear/ Non Linear/ Polynomial/ Logistic Regression. Moreover, K-Nearest Neighbors, Support Vector Machines (SVMs), Random Forests and Decision Trees. Importantly, advanced supervised methods are included, such as ensembled techniques and Neural Networks (NNs). Furthermore, unsupervised learning methods are discussed, including Kmeans, Hierarchical Clustering and Latent Variable Models, while the authors provide a brief theoretical overview of Reinforcement Learning (RL).

Finally, through some strikingly historical examples, the writers demonstrate the value of the proper use of Machine Learning in state-of-art projects. It is essential to notice that the proper usage of ML is defined in three axes.

- Bias
- Authority

- Effectiveness

An indicative example of the "Effectiveness ax" is the case of Thalidomide. After World War II a company, located in West German, named Chemie Grünenthal was searching and working in order to full fill the demand for further antibiotics. While researching, they found a painkiller medicine, called Thalidomide, that had also anti-emetic effects. In 1957, the drug was strongly advertised, for soothing morning nausea of pregnant women. Over the course of the drug's marketing campaign, teratogenic birth abnormalities were detected in thousands of newborns. This led to the gradual removal of the drug from the market. The inconsiderate action, of selling this substance to pregnant women only without the vital extensive background research, was not recognized, from the German government as negligent homicide or injury. The excuse of the company was that they were aware of the "nothing-crosses-the-placental-barrier"; as such the drug could not influence the infant. However, it is commonly known, since 1957, that alcohol penetrates through the placental barrier, which provides strong indication that any other drug could, potentially, cross it. This case shows a harmful fallacy presented as proof of the mindset "it just works". A model or algorithm does not automatically qualify as "excellent" just because it "works" in the sense of producing a desirable result, such as easing temporary symptoms like financial limitations. It is important to distinguish between "it works" and "it is good," as doing so would imply that additional critical analysis is not necessary.

### *B. A Comparison of Neural Networks and Support Vector Machines Algorithm Performance in Human Resource Analytics*

In this paper the authors, namely Hannes Draxl and Ryan Nazareth, highlight the fact that attrition is not only a time-consuming but also an expensive process. Considering that there are a plethora of factors which play a significant role into the decision of an employee to quit the company, the writers attempt to predict this decision, in advance, utilizing robust ML algorithms. Specifically, this study's primary goal is to evaluate and compare the effectiveness of the Multi-Layer Perceptron (MLP) algorithm and in contrast to Support Vector Machine (SVM) classifier.

The target metric is whether an employee is going to quit. The name of this variable is "Left" and is divided in two categories:

- stay
- leave

After exploring the columns of the dataset and indicating the target variable the authors implement Explanatory Analysis. In particular, they use *Heat map* in order to visualize the correlation between the features. They observe that the "Left" variable (target variable) has a strong correlation (-0.5) with the "satisfaction_level" feature. So, we conclude that the decision of quitting is affected by their degree of contentedness. Another interesting way of plotting the correlation is used called the *violin plot*. From this plot, it is obvious that the salary feature in combination with the promotion_last_5years can increase the prediction power of the model.

For the training and testing process they split the data into 70% training and 30% test. Then, they used 5-fold stratified cross validation with embedded grid search for hyper parameter tuning. The models with the best HP were retrained using the entire training set after grid search. In order to determine how many samples the models can correctly categorize as being on the left out of all samples that correspond to "left," they relied on a confusion matrix and in particular the recall score (the true positive rate).

In conclusion, SVM and MLP performed well in forecasting whether or not a customer will quit the organization, and as a result, is of great importance for human resource departments. SVM and MLP demonstrated both their strengths and shortcomings during the analysis; they are both strong algorithms but take a lot of computational power and time to tweak and assess. Results showed that SVM performed better than MLP in terms of test accuracy overall (F1 score), while MLP had a slightly greater recall rate than SVM, which is consistent with their original premise.

### *C. IBM HR Analytics Employee Attrition*

In this research, that was initiated by the IBM company, the main purpose is to create a profile (find features) of the employees that are going to quit their job. In order to achieve this, they made a data-set which contains demographic and work related information. Some names of the columns are the follow:

- Age
- Education
- Department
- MonthlyIncome
- EnvironmentSatisfaction
- JobInvolvement

Subsequently, they performed Statistical Analysis and specifically created figures which demonstrate relations between the features. For instance, in the plot where is depicted the "Attrition" with the "MaritalStatus" columns, we discern a pattern where the single employees have bigger attrition than those who are married and married workers have greater attrition in comparison with the divorced. Furthermore, another significant inference can be made from the "Department" and "Attrition" plot. We can clearly see that the Sales department has the highest attrition and the HR department has the lowest.

Following, a Decision Tree algorithm is implemented as a way to find and distinguish the *important features*. The weighted decrease in node impurity divided by the likelihood of reaching that node is used to determine a feature's relevance. The node probability can be computed by dividing the total number of samples by the amount of samples that reach the node. The feature is more significant the higher the value. After plotting the importance and the features we notice that the most important columns are the:

- MonthlyIncome
- Age
- OverTime

- DistanceFromHome
- TotalWorkingYears
- YearsWithCurrManager
- NumCompaniesWorked

After finding the significant variables, the authors, perform the K Means algorithm with 2 clusters. The conclusions are that Cluster 0 characteristics are:

- lower attrition rate
- more senior jobs
- higher salary
- greater number of companies worked for
- older age
- fewer single persons
- more YearsAtCompany
- fewer YearsSinceLastPromotion

By the time this research was completed the writers highlight 3 important findings.

*1) Finding 1:* Despite having a higher rate of promotion and salary than other non-managerial positions, the human resources department has a high attrition rate.

*2) Finding 2:* People who worked in two to four companies over the course of their careers are less likely to depart. After working for six firms, female attrition is significantly lower than male attrition.

*3) Finding 3:* Compared to other work levels where doctors nearly always have the lowest attrition rate, Job Level 3 has the greatest attrition rate for medical professionals. One explanation could be that it takes more time for doctors to reach job level 4.

Finally, it is worth mentioning some managerial advice the authors offer. To begin with, they suggest that the business should examine human resource positions more thoroughly to see which aspects of the job are unsatisfactory to employees. It is very advised to have one-on-one conversations regularly. Moreover, they recommend that males who attended more than five companies should still be given attention, even though the corporation is not required to concern much about those who worked for two to four different businesses. Lastly, they draw attention to the fact that doctors are taking longer to go from level 3 to level 4 jobs and are less pleased with their level 3. To ensure employees are adequately compensated, it is advised to carefully analyze the performance evaluation method.

### *D. Walking assets: The cost of losing an employee*

In this work the author Juan Martin Carriquiry is doing a research on business's expenditures associated with voluntary employee turnover. To calculate the expenses of engineer turnover, evidence from a knowledge-intensive manufacturing company is specifically employed. Data was gathered from company documents, management and HR staff interviews. According to the research, staff turnover is a costly phenomena. The productivity loss is by far the biggest single financial burden for the business. This suggests that turnover is particularly expensive for jobs involving difficult tasks and lengthy learning curves.

The analysis of high attrition costs for the company should help with resource distribution in an effective manner. How much funding should go toward addressing sub-optimal turnover rates is a constant concern for managers, especially in the area of HR decision-making.

The writer implements a model which takes into consideration the expenses of decreased productivity caused by turnover in order to address the issue of the absence of precise information on the growth of employee productivity.

The addition of literature is that we may get a much fuller understanding of the actual effects on the firm by factoring learning processes into the evaluation of staff turnover. These results support Waldman et al(2004) .'s strategy of include learning curves at the individual level to estimate turnover costs and demonstrate that this methodology might be applied to other industries. This method should be used by practitioners to determine the actual financial loss caused by staff turnover and be included in financial statements and balances.

Juan Martin Carriquiry refers to tree different opinions regarding the fact to what extent employee behavior that results in attrition is bad for the business. The first one, mentioned in Baron et al.(2001) [5], considers employee turnover a negative case. Specifically, staff turnover is considered as an event that disrupts organizational patterns and hence poses a performance risk for the company. The second aspect expressed in another published work (Dalton and Todor (1979), Staw (1980), Dalton et al. (1981) [6] [7] [8]) states that staff turnover is an overestimated fact. They contend that because staff are generally simple to replace, turnover only slightly lowers output. Finally, according to the (Abelson and Baysinger (1984), Glebbeek and Bax (2004) [9]) work the real question is "what percentage of turnover is ideal for the company". Retention techniques have expenses associated with them, just as turnover strategies have costs associated with it. According to this cost-benefit analysis, the "ideal" rate of turnover should occur when the marginal costs of turnover and retention are equal.

This paper suggests that no business should strive for a zero turnover rate. According to them, management should aim to achieve a "optimal" turnover rate, which is defined as the point at which the net costs of attrition are equal to zero. Further efforts to keep employees would be suboptimal and financially ineffective at that point since the costs of retention strategies would equal the expenses of employee retention. This work's main topic is the economic effect of attrition on the company.

By utilizing a learning curve technique, like Waldman et al. $(2004)^2$ [10] did in his examination of turnover in the health care industry, this paper will contribute to the field. This approach models learning as a continuous, non-linear process rather than a discrete one.

The full accounting of staff turnover expenses contains:

- **Departure of an employee** which includes administrative expenses, departure interviews conducted by the human manager and a representative from human resources, the manager's restructuring of the tasks, and IT expenses. When an employee retires, supervisors may need to rearrange their responsibilities, which might take time. If

a temporary successor is needed for the retiring employee, the business may incur additional expenses. If the burden is distributed among the remaining staff, there may be increased overtime costs and lower productivity.
- Recruiting
- Educating and instructing the new employee
- Costs of decreased productivity as a result of the new worker's poor productivity

The early productivity of new personnel is typically lower than that of preceding workers. According to the learning curve idea, the more practice you get, the more proficient you can be at the subject. The type of work will determine the form of the learning curve. The author employs a variant of the well-known parabolic curve model (Hackett, 1983) [11], which is structurally related to the accumulative learning methods (Mazur and Hastie, 1978). It is not universally accepted that learning is a smooth, monotonic process in which learning gradually declines; in reality, learning curves with a wide range of forms have been seen (Mazur and Hastie, 1978). Their model presupposes that learning processes occur at steady rate.

Output loss:

$$y_i = 1 - \frac{1}{ax_i + Exp} \tag{1}$$

Then becomes:

$$y_i = -\frac{1}{ax_i + Exp} \tag{2}$$

Where:

- x: time function
- a: learning ability
- Exp: previous experience

and

$$Y_i = W_i * \int_i^{i+1} F(x)\,dx \tag{3}$$

Where:

- $Y_i$: productivity loss
- $W_i$: salary for a specific period of time

After, he expands the model to account for various payments throughout numerous time periods.

$$Y(i,N) = \sum_{i=0}^{i=N} W_i * \int_i^N F(x)\,dx \tag{4}$$

***Data***:

The HR department submitted firm documents, including information on remuneration, recruitment, commencement, and training expenses, from which attrition figures were produced. Manager statements were used to assess the expenses of additional training and decreased productivity. There was no information available about the productivity of the departing staff. To acquire information that would suggest a skewed attrition of lesser or better employees, managers were indeed asked what the effectiveness level of the engineers who left in the past year compared to those who remained was. In order to calculate the coeffcients for the modeling of learning curves, managers were furthermore requested to judge the attribution of workers at various time periods.

All of the engineers who departed the business between 2010 and 2011 make up the sample. 37 of the 340 engineers departed from the company in the past two years. 24 of these workers left the company voluntarily, 12 completed their short-term contracts, and 1 was fired.

The variables consist of the *departure, hiring, training costs* and the *productivity* which was calculated from the interviews conducted by the managers.

***Results –Total costs of employee resignation***:

| Item | Cost per employee | Total cost |
|---|---|---|
| Departure | 652 | 14,345 |
| Hiring | 2,546 | 56,003 |
| Training | 8,097 | 178,134 |
| Loss productivity | 30,254 | 665,592 |
| Total costs | 41,549 | 914,074 |

Fig. 3: Total costs

The statistical analysis's findings indicate that, overall, it costs $41,549 to replace an engineer. The expense of decreased productivity, which accounted for over 70% of the overall costs, was by far the greatest factor. The research also demonstrates that, once salary differences are accounted, the value of the productivity loss after 4-5 years is quite comparable regardless of whether the new recruit has no experience or much of it. As anticipated, the engineers leaving the firm on general were younger and had worked for the business significantly less time than those staying.

Primarily, a significant portion of the costs of turnover are attributable to the learning process in the highly qualified roles taken into account. This is very pertinent since it emphasizes the financial costs to the company of training new staff.

According to this study, a variety of elements will affect the expenses of staff turnover for the company. First, the firm's recruiting procedures, which may be influenced by cultural and professional conventions. Second, the traits of the new employee, especially if they have prior experience in the field. The expenditures of employee turnover in this scenario do not seem to be much impacted by experience. Third, the technology and expertise needed for the job. These factors demonstrate that the expenses related with training new personnel account for the majority of the expense in the process of turnover.

## III. System Architecture

This section describes the process of developing an intelligent system that is useful for the Human Resources Management department. Specifically, a real dataset is studied and robust models are developed to forecast the probability of an employee to search for a new job or stay in the company. In subsection III-A, an overview of the dataset is provided.

The utilized models are described in subsection III-C, while III-B discusses about the handling of the data.

### A. Data

The analyzed dataset is the *"HR Analytics: Job Change of Data Scientists"* [13]. This dataset is designed to find the factors that make a person to leave current job by analyzing key elements, such as the current credentials, demographics and experience data. It consists of 14 features (columns) and 19159 rows. The features are:

- ***enrollee_id*** : Unique ID for candidate
- ***city***: City code
- ***city_development_index***: Developement index of the city (scaled)
- ***gender***: Gender of candidate
- ***relevent_experience***: Relevant experience of candidate
- ***enrolled_university***: Type of University course enrolled if any
- ***education_level***: Education level of candidate
- ***major_discipline***: Education major discipline of candidate
- ***experience***: Candidate total experience in years
- ***company_size***: No of employees in current employer's company
- ***company_type*** : Type of current employer
- ***lastnewjob***: Difference in years between previous job and current job
- ***training_hours***: training hours completed
- ***target***: 0 – Not looking for job change, 1 – Looking for a job change

### B. Data Management

The appropriate data management and pre-processing is of high importance given that the dataset is flawed in a lot of ways:

- The dataset is imbalanced.
- Most features are categorical (Nominal, Ordinal, Binary), some with high cardinality.
- Missing imputation are a part of the pipeline as well.

To deal with these issues, we follow several important steps. At first, we use the *Label Encoding* technique to transform the labels into numeric form (machine-readable form). This is an essential step as the ML algorithms can only handle data in numeric form. Importantly, we use mapping dictionaries for ordinal features to provide a mapping that takes care of the different levels of a feature. For instance, the *education_level* feature has a lot of values (phd, primary, masters, ...) that should be ordered appropriately. The proper way to do it is with the following mapping:

- 'Primary School' : 0,
- 'High School' : 1,
- 'Graduate' : 2,
- 'Masters' : 3,
- 'Phd' : 4

Following that, we use k-Nearest Neighbors (k-NN) imputer implementation to fill the missing values in the dataset. In specific, a sample with missing values is selected and the nearest neighbours are found utilising some kind of distance metric; for instance, the Euclidean distance. Then, the mean of all nearest neighbours is calculated and the missing value is filled up.

Furthermore, we utilize the Synthetic Minority Oversampling TEchnique (SMOTE) to increase the data in a balanced manner alleviating the "imbalanced dataset" issue. This is done by oversampling the examples in the minority class synthesizing new examples. SMOTE choses examples close in the feature space, creating a line among them and creating a new sample at a point along that line. This way, we develop a balanced dataset that will be the proper base to build powerful models that will generalize well avoiding the overfitting issue.

### C. Algorithms

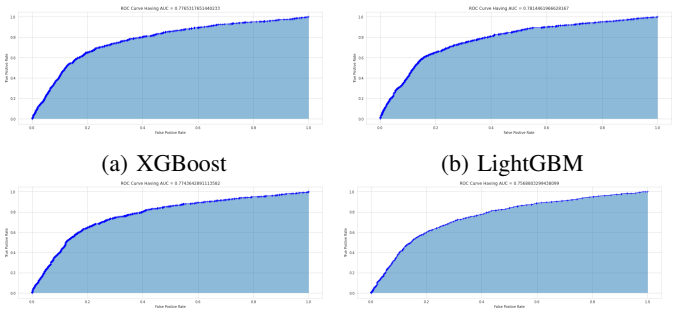
The utilized machine learning algorithms are ensemble techniques, namely XGBoost, LightGBM, CatBoost and Random Forest. They are all ensemble techniques and the first three of them apply the boosting method. In ML, boosting is an ensemble meta-algorithm for minimizing bias and variance in supervised learning. Boosting converts weak learners to strong ones. A weak learner is a classifier that classifies data correctly few times. In contrast, a strong learner is a classifier that finds the right classification almost every time. On the other hand, Random Forest applies the Bootstrap aggregating (bagging) technique. Bagging is an ML ensemble meta-algorithm that enhances the stability and accuracy of the models used. It moreover minimizes variance and helps to avoid overfitting. It is usually applied to decision tree methods.

## IV. Evaluation

To evaluate our models we utilize the ROC curve and the AUC score. Receiver Operating Characteristics (ROC) curve is a plot presenting the efficiency of a classification model at all thresholds. This curve is created by plotting two parameters, the true positive rate (TPR) against the false positive rate (FPR). Area Under Curve (AUC) score shows the degree or measure of separability. A model with higher AUC is better at forecasting True Positives and True Negatives. AUC score calculates the total area below the ROC curve. AUC is scale invariant and also threshold invariant. In probability terminology, AUC score equals to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. To find the best possible configuration of every model, we extensively experimented with the different hyper-parameters (tuning) using Bayesian optimization. Bayesian Optimization is employed in applied ML to tune the hyper-parameters of a given well-performing model on a validation set. Below, we provide the evaluation results of the best configuration for every model.

### A. XGBoost

XGBoost performed remarkably well with the approrpiate hyper-parameter tuning. Specifically, the optimal hyper-parameter configuration is:



(a) XGBoost (b) LightGBM

(c) CatBoost (d) Random Forest

Fig. 4: ROC Curves of the ensemble models on validation (test) data.

- $n\_estimators$: 1427,
- $min\_child\_weight$: 14.5,
- $gamma$: 2.09,
- $colsample\_bytree$: 0.97

It managed to achieve a training AUC Score of 0.93 and a test AUC score of 0.78. The respective ROC curve for the testing data is illustrated in figure 4a.

### B. LightGBM

LightGBM has a great performance with the approrpiate hyper-parameter tuning. Specifically, the optimal hyper-parameter configuration is:

- $n\_estimators$: 100,
- $min\_child\_weight$: 1.05

It managed to achieve a training AUC Score of 0.95 and a test AUC score of 0.78. The respective ROC curve for the testing data is illustrated in figure 4b.

### C. CatBoost

CatBoost was also very efficient with the proper hyper-parameter tuning. Specifically, the optimal hyper-parameter configuration is with $n\_estimators$ equal to 135. It achieves a training AUC Score of 0.94 and a test AUC score of 0.77. The respective ROC curve of test data is shown in figure 4c.

### D. Random Forest

Random Forest performed excellently using the right hyper-parameters. Specifically, the optimal hyper-parameter configuration is with $n\_estimators$ equal to 174. It achieves a training AUC Score of 0.99 and a test AUC score of 0.76. The respective ROC curve of test data is shown in figure 4d.

### E. Summary

The results are summarized in figure 5 depicting the models' great generalization performance. Generally, the models avoided over-fitting and reached excellent classification scores. Thus, we recommend these models for the studied problem as they combine high accuracy with little training time.

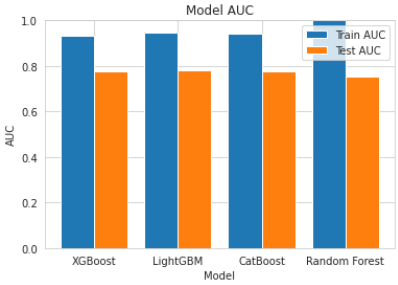



Fig. 5: Evaluation Summary

## V. Conclusion

In this work, we investigate and summarize the literature on Human Resources (HR) Analytics by presenting three of the most influential publications and a valuable book. Further, we delve into a real dataset and develop a sophisticated AI system that is able to make forecasting about the probability of an employee to leave or stay in the company. Next, we proposed and evaluated four ensemble techniques that performed excellently. In the future, we look forward to extend this work by gathering more data, analyzing more dataset and investigating more machine learning and deep learning algorithms. We will also investigate more topics in HR Analytics to broaden our views and find efficient solutions.

## References


[1] AI & Automation, [online] https://www.gsma.com/futurenetworks/wiki/ai-automation-anoverview/?fbclid=IwAR1gv-lMewcrYvXYFrLwW_120E5oreF9zjzGX9CCJVNuaFlPcPleBX5XTpQ . Access Date: 5/07/2022

[2] How HR Analytics Are Changing Business, [online] https://lesley.edu/article/howhranalyticsarechangingbusiness . Access Date: 17/07/2022

[3] Human Resource Management Review, Volume 32, Issue 2, June 2022, 100795 Human Resource Management Review Volume 32, Issue 2, June 2022, 100795, Human resources analytics: A systematization of research topics and directions for future research

[4] Deductive vs Inductive Reasoning, Inductive vs. Deductive Research Approach (with Examples), [online] https://www.scribbr.com/methodology/inductive-deductive-reasoning/

[5] Baron, J., M. Hannan, and M. Burton (2001). Labor pains: Change in organizational models and employee turnover in young, high-tech rms. American Journal of Sociology 106 (4), 960-1012.

[6] Dalton, D. and W. Todor (1979). Turnover turned over: An expanded and positive perspective. Academy of Management Review 4 (2), 225-235.

[7] Staw, B. (1980). The consequences of turnover. Journal of Occupational Behaviour 1 (4), 253-273.

[8] Dalton, D., W. Todor, and D. Krackhardt (1982). Turnover overstated: The functional taxonomy. Academy of management Review 7 (1), 117-123.

[9] Abelson, M. and B. Baysinger (1984). Optimal and dysfunctional turnover: Toward an organizational level model. Academy of Management Review 9 (2), 331-341.

[10] Waldman, J., F. Kelly, S. Arora, and H. Smith (2004). The shocking cost of turnover in health care. Health Care Management Review 29 (1),2.

[11] Mazur, J. and R. Hastie (1978). Learning as accumulation: A reexamination of the learning curve. Psychological Bulletin 85 (6), 1256.

[12] Juan Martin Carriquiry. Walking assets: The cost of losing an employee. Aalborg University January 1, 2013.

[13] HR Analytics: Job Change of Data Scientists, [online] https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists?resource=download&select=aug_train.csv . Access Date: 5/07/2022.

[14] Christopher M. Rosett Austin Hagerty. Introducing HR Analytics with Machine Learning.