UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA CIVILE DELL'ENERGIA DEI SISTEMI DEL TERRITORIO E DELLE COSTRUZIONI

RELAZIONE PER IL CONSEGUIMENTO DELLA LAUREA MAGISTRALE IN INGEGNERIA GESTIONALE

HR Advanced Analytics: The Employee Attrition Problem. A Case Study

SINTESI

RELATORI

Prof. Antonella Martini
Dipartimento di Ingegneria dell'Energia dei Sistemi, del Territorio e delle Costruzioni

Ing. Lorenzo Baiocco
ELIS Consulting & Labs

CANDIDATO

Francesco Ghezzi
francesco.ghezzi2@gmail.com

Sessione di Laurea Magistrale del 17/02/2021
Anno Accademico 2019/2020
Consultazione NON consentita

HR Advanced Analytics: The Employee Attrition Problem. A Case Study

Francesco Ghezzi

Sommario

Il turnover dei dipendenti è un grave problema per tutte le imprese di qualsiasi settore. Le cause del fenomeno sono molteplici e difficili da individuare, poiché possono variare da fattori macro-economici a ragioni personali del singolo. Inoltre, i costi associati sono elevati, in particolar modo quando lasciano l'azienda dipendenti di elevata esperienza. Tutto ciò fa sì che elevati tassi di turnover siano insostenibili per un'impresa. Questa tesi affronta il problema del turnover dei dipendenti facendo uso di predictive analytics: andando oltre la semplice descrizione del perché un dipendente lasci la propria impresa, l'obiettivo è quello di (1) sviluppare un modello predittivo dei dipendenti a rischio di turnover e (2) di implementarlo in un caso studio. A questo scopo, sono state impiegate varie tecniche di machine learning, utilizzate in letteratura nell'ambito della predizione del turnover dei dipendenti e ne sono state valutate le performance, così da selezionare la più performante. Il caso studio condotto nasce da un progetto di 4 mesi, realizzato all'interno del programma Junior Consulting, presso ELIS Consulting & Labs a Roma.

Abstract

Employee attrition is a severe issue for all companies in every industry. The causes of this phenomenon are multiples and difficult to detect, ranging from macro-economical factors to personal concerns of the single employee. Moreover, significant costs are associated with attrition, in particular when skilled and experienced employees leave. It results that high turnover rates are disruptive for a firm. This thesis addresses this problem through predictive analytics: going beyond the simple description of why an employee leaves a company, the purpose is to develop a comprehensive framework in predicting which employees are at risk of leaving and to implement it on the proposed case study. In order to do this, various machine learning techniques used in literature to predict employee attrition have been implemented and their performance assessed to select the best among them. The case study originates from a 4-months project carried out within the Junior Consulting program, at ELIS Consulting & Labs in Rome.

1 Context and Goals

This thesis arises from a project which took place in ELIS Consulting&Labs in Rome. The project was commissioned by a large international company, that operates in the energy industry. The study lasted four months and I have directly collaborated in its realization in all the phases. The project originates from the client's need to define a model that allows for a thorough understanding of the relationships between the main variables available in its information systems and the employee attrition phenomenon.

Employee attrition is defined as a *"reduction in the number of people who work for an organization that is achieved by not replacing those people who leave"* [1]. Attrition can be either voluntary or involuntary: this thesis' focus is on voluntary attrition as it is predictable through employee data, while involuntary attrition is led by the firm and its internal logics. Attrition can originate at multiple levels and understanding why an employee is leaving is quite a complicated issue: it can be related to *Job-Related Factors*, *Organizational Factors* and *Personal Factors*.

Employee attrition is a severe concern for firms because of the heavy costs involved: it is estimated that, depending on experience and skill levels, the cost of loosing an employee ranges between 16% to 213% of their annual salary[2]. The costs are both direct - such as replacement and training costs - and indirect -such as knowledge leak, loss of productivity and reduced morale in the working environment.

In the HR field structured data driven approaches are not so frequently implemented, because HR databases are often very noisy: information is collected in multiple heterogenous ways and a lot of data are sensitive, making companies unwilling to share it. On the contrary, data driven approaches can be extremely useful in analyzing employee's data. In particular, automated methods, like machine learning ones, can be used to solve complex problems, as the one addressed by this project. In this context, an application of these techniques to extract employee attrition HR analytics is a promising perspective for every company, in particular for a large one with heterogeneous workforce and great availability of HR data.

2 Project's Overview

Starting from the client's requirement of understanding which employees are at risk of leaving, the problem was formally defined as a machine learning binary classification task, where the two classes are *leaving* and *remaining*.

The project was organized according to the Agile methodology, and the activities were grouped in three main phases: *state of the art analysis, benchmark analysis,* and *predictive analysis*. In Table 1 phases, methodological steps, main outcomes and the thesis paragraphs are reported.

[1] Cambridge Advanced Learner's Dictionary and Thesaurus, Cambridge University Press, 2020.

[2] Boushey, H., & Glynn, S. J. (2012). There are significant business costs to replacing employees. Center for American Progress, 16, 1–9.

Phase	Sub-Phase	Methodological Steps	Outcomes	§
State of the Art Analysis		-Query design -Search of articles on Scopus -Careful analysis and selection of the more coherent to the scope	-Collection of 79 papers and final selection of 19 -Definition of the main characteristics of the employee attrition problem	3
Benchmark Analysis	Salary Benchmark	-Data collection -Data organization in clusters and time series -Data exploration	-20 publications collected -Overview of the salaries market in Italy and in Energy industry from 2015 to 2019	4.1.1, 4.2.1
	Turnover Benchmark	-Data collection -Data organization in time series -Data exploration	-Overview of entry, exit, balance and turnover rates in Italy and in Energy industry from 2015 to 2020	4.1.2, 4.2.2
Predictive Analysis	Data Collection	-Collection from the client's information systems -Linking key creation	-5 databases collected	5.1.1, 5.2.2
	Data Preparation	-Data Exploration -Data Cleaning -Feature Engineering -Final preparation	-Creation of a single dataset from the initial 5 -Improvement in Data Quality -13 features engineered	5.1.2, 5.2.2
	Data Description	-Use of graphical charts and bars -Correlation analysis	-Highlighting of interesting relationships and correlations with the target variable	5.1.3, 5.2.3
	Data Prediction	-Models implementation -Performance Metric Selection -Training and Testing -Performance evaluation	-6 models implemented -3 different datasets used -3 metrics selected -Comparison of performance among different models	5.1.4, 5.2.4
	Final Model Selection	-Selection of the best performing model and results presentation	-CART on the feature-reduced dataset -Recall=1 -F1 Score=0,98 -ROC-AUC=0,994	5.1.5, 5.2.5

Table 1: Project's Phases, Sub-Phases, Methodological steps and Results

3 State of the Art Analysis

The State of the Art Analysis is aimed at discovering *what is known* behind the employee attrition prediction literature and the most recent findings. Since the '70s of the XX century, lots of works concentrated on understanding the antecedents of the attrition behavior, analyzing and testing correlations with multiple factors and building the more common turnover paths. In recent years, a branch of the research on employee attrition diverges from the classical descriptive investigation towards a predictive focus, in which machine learning techniques are employed.

To analyze this field of studies, Scopus database was used as source and a query design process was undertaken in order to select a set of papers needed to carry out a proper Literature Review (Table 2).

Activity	Methodology and Results
Query Design and Papers Collection	Research criteria: -Database: Scopus -Query: (“employee attrition” OR ”employee churn” OR ”employee turnover”) AND(”prediction”OR”predict”OR”predictive”OR”analytics”) -Subject Area: Every Area -Location: Titles, Keywords, Abstracts -Time Horizon: No Limits This query led to 237 results. Among them, 79 were collected as related to the topic of the research
Papers Selection and Analysis	Papers were carefully read, and 19 were finally selected as of most interest with respect to the project scope. For each of them, whenever possible, the Scope/Context of the paper, the datasets used, the algorithms implemented, the metrics chosen and the main insights, such as relevant variables and best performing algorithms, were highlighted.

Table 2: Methodology and Results of Literature Review activities

3.1 Results

The state of the art analysis highlighted some interesting and useful insights about the employee attrition prediction problem:

- HR datasets are frequently very noisy; moreover, every company has its own particular environment, internal and external, which the more distant the companies (in terms of culture, industry, country) the more different it is. This affects the reliability of any general conclusion, depending too much on the dataset used;
- Class imbalance is an evergreen concern: churners are always a minor percentage of the total employees (on average between 10% and 15%). That means re-sampling[3] is always necessary and, in terms of evaluation metrics, accuracy is not sufficient to assess models’ performance because it is biased by the true negative ratio. F1 Score and ROC-AUC[4] can be considered more reliable;
- Random Forest (RF) is recognized as the best performing algorithm when it comes to employee attrition prediction. However, two facts should be noted:
    1. Whenever XGBoost is used, it outperforms even RF thanks to a better dealing with overfitting;
    2. Classification Trees such as C5.0 and CART frequently follow, in terms of performance, RF in a close range but at the same time they are far less complex, indeed some authors identify them as a good compromise for practical implementations.

[3]Re-sampling is a dataset-transformation which is performed by over-sampling the minority class or by under-sampling the majority class

[4]Area Under the Curve of the Receiver Operating Characteristic curve

Anyway, all the best algorithms are decision trees because RF and XGBoost are ensembles of decision trees;

- The most relevant features in predicting employee attrition are Salary, Seniority[5], and Overtime[6].

4 Benchmark Analysis

The Benchmark Analysis phase is divided in the following sub-phases: Salary Benchmark, and Turnover Benchmark.

4.1 Salary Benchmark

The objective of Salary Benchmark is making a comparison between what the market pro vides, in terms of salary, for a certain position and what the client company does, throughout the time. Thus, the following methodology was formulated: collection of data from the major HR provider & consulting firms (Spring Professional, Hays, Job Pricing), ranging from 2015 to 2019; selection and export of the relevant information; data organization: cluster- ing by qualification (labourers, clerks, managers, executives), by business area (Production, R&D, IT, Legal, HR..) and by educational degree (Graduated/Not Graduated), creation of time series from 2015 to 2019, division between average Italian data and Energy industry- specific data; validation comparing different sources; data exploration.

4.1.1 Results

A total of 20 publications were collected and studied. As a first step, data about Italian and Energy industry yearly salaries per qualification were selected and analyzed in order to draw relevant insights. In Figure 1 the trends throughout the years of these variables are plotted. The average salary level of Energy industry is clearly above Italian one, even if for executives it happens only from 2019 on. Then, salaries were studied from the points of view of experience, age and business areas. It was highlighted that both experience and age are positively correlated with average

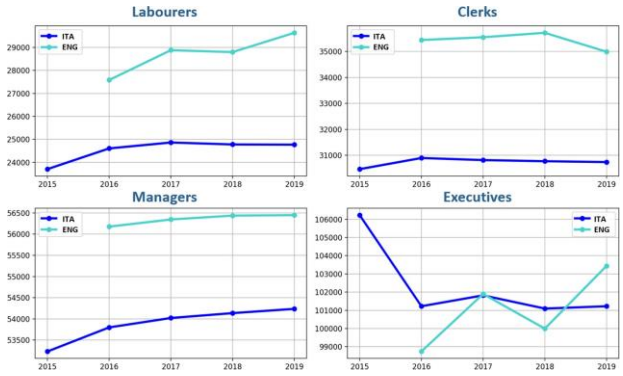


Figure 1: Comparison of Italy and Energy industry average RAL, from 2015 to 2019

[5]Length of service/years at the company
[6]Hours worked after the normal schedule

salaries and in particular there is a marked hike in salary when experience becomes greater than 5 years. Passing to business areas, it was highlighted that both in Italy and Energy industry the Sales area provides, on average, the higher salaries, both for clerks and for managers. On the contrary, clerks in Production and managers in IT has the lowerones.

4.2 Turnover Benchmark

As a second step of the Benchmark Analysis, the focus was set on job movements and turnover rates. This sub-phase is aimed at analyzing the manifestation of the employee attrition phenomenon in Italy and in Energy industry, studying average entry and exit rates of employees, turnover rates and balances. In order to perform this analysis, the following process was set up: collection and export of raw data (monthly entry and exit rates from 2015 to July 2020) from ISTAT[7]; data organization in monthly time series; calculation of turnover and total balance rates; data exploration.

4.2.1 Results

Data extracted covered the ATECO classes B_S_NO-O-P[8] and D[9]. The time series of entry and exit rates for both these classes can be seen in Figure 2. The figure clearly shows that Energy industry has lower entry and exit rates. Balance rate in a period of time t_0 was

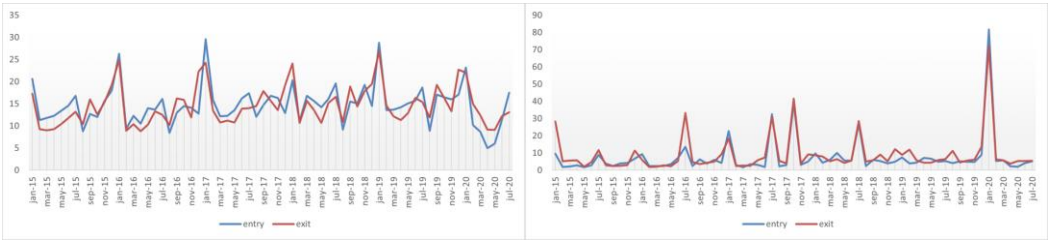


Figure 2: Monthly Entry and Exit Rates of employees in Italy (on the left) and in Energy industry (on the right), from 2015 to July 2020

calculated as $B_{t0} = INRATE_{t0} - OUTRATE_{t0}$, while turnover rate at t_0 was calculated as $T_{t0} = INRATE_{t0} + OUTRATE_{t0}$. The analysis of these rates over time showed that turnover in Energy industry is significantly lower than Italian market one (as it can be seen from Figure 2), but at the same time in the last 5 years the Energy industry has suffered a continuous decline in the total number of workers: Labourers fell by 15%, while the fall in Managers and Clerks was around -7%.

[7]http://dati.istat.it/Index.aspx?QueryId=26682
[8]Manufacturing+Services
[9]Energy/Utilities

5 Predictive Analysis

The Predictive Analysis phase is divided in the following sub-phases: Data Collection, Data Preparation, Data Description, Data Prediction, and Final Model Selection.

5.1 Data Collection

The Data Collection phase is the first step in every data science pipeline. Data was collected from the information systems of the client. A total of 5 databases were available, each with different features and/or time horizon. The databases contained not only employee-related data, but also related to job movements and remuneration events. The creation of a key to connect the different databases has been the first and fundamental task of this phase; after the necessary anonymization of employees in the dataset, the key has been identified in the feature *ID*, that is the identification number of a single employee.
The databases collected are: "DATI_DIPENDENTI_ATTIVI"; "STORICO_MOVIMENTAZIONI_2020"; "STORICO_MOVIMENTAZIONI"; "STORICO_RETRIBUTIVO_2020"; "STORICO_RETRIBUTIVO".

5.2 Data Preparation

The Data Preparation phase is divided in the following sub-tasks: (a) Data Exploration, (b) Data Cleaning, (c) Feature Engineering, (d) Final Preparation. The objective of this phase is to obtain a unique dataset, provided with quality data and all the information needed for the following phases.
(a) Data Exploration is aimed at understanding the macro-characteristics of the datasets. In this phase, for each dataset, the following characteristics were analyzed: number of records, number and types of features, number of missing values, number of different IDs.
(b) Data Cleaning is the activity of manipulating the data in order to make it available for any further analysis. This is a necessary step because incorrect or inconsistent data leads to false conclusions[10]. For each dataset, the following tasks were performed: Removal of irrelevant data; Removal of duplicates; Handling of missing values; Types conversion; Column names standardization; Values-formats standardization.
(c) Feature Engineering follows the cleaning phase and it aims at building new features on the cleaned dataset in order to get more information. The feature engineering task was carried out through the detection of valuable information to engineer, considering both the most relevant employee attrition antecedents, as extracted from literature, and the available raw information; the definition of the operations needed to build the features; the mapping of the features in the dataset "DATI_DIPENDENTI_ATTIVI" for every available ID.
Lastly, (d) Final Preparation concerns the last preparatory tasks performed in order to have the dataset ready for the prediction phase. The following activities were required: removal of remaining missing values; one-hot encoding[11] of categorical features; removal of redun-

[10] Omar Elgabry, The Ultimate Guide to Data Cleaning
[11] Method used to convert categorical features in quantitative ones

-dant features; split of the dataset between independent variables and target variable.

5.2.1 Results

Datasets collected were characterized as below:

- ”DATI_DIPENDENTI_ATTIVI”: 2078 rows, 13 features, 2077 different IDs;
- ”STORICO_MOVIMENTAZIONI_2020”: 875 rows, 27 features, 636 different IDs;
- ”STORICO_MOVIMENTAZIONI”: 53948 rows, 13 features, 2053 different IDs;
- ”STORICO_RETRIBUTIVO_2020”: 1107 rows, 7 features, 1033 different IDs;
- ”STORICO_RETRIBUTIVO”: 32084 rows, 7 features, 2047 different IDs.

After Cleaning phase, data quality was ensured and inconsistencies among different datasets resolved. That made possible to merge movimentation datasets (”STORICO_MOVIMENTAZIONI_2020” and ”STORICO_MOVIMENTAZIONI”) in a unique dataset called ”MOVIMENTATIONS”, as well as remuneration datasets (”STORICO_RETRIBUTIVO_2020” and ”STORICO_RETRIBUTIVO”) in a unique dataset called ”REMUNERATIONS”.
A total of 13 features were engineered and mapped into “DATI_DIPENDENTI_ATTIVI”, taking information from “MOVIMENTATIONS” and “REMUNERATIONS” datasets and from the results of the Benchmark Analysis phase. Among them, they were engineered also the target variable of the prediction models, which is the feature *Attrition*, and the fea- ture *Salary Gap*. The first one is a binary variable whose possible values are ’leaving’ and ’remaining’. Since “DATI_DIPENDENTI_ATTIVI” is updated at the end of 2019, each employee was considered as belonging to ’leaving’ class if linked an exit movement during 2020 in the dataset “MOVIMENTATIONS”, otherwise as belonging to ’remaining’ class. The second one was calculated as the difference between the actual salary of an employee and the average one[12] provided by the market, given the same qualification and business area.
After removal of missing values, a total of 2034 rows were left. Moreover, the features one-hot encoded were: *Gender, Type of Education Title, Society, Business Area, Attrition* and *Qualification*. This transformation produced 28 binary variable (0-1), one for every possible value of each categorical variable. In particular, considering the target variable, ’leaving’ class was encoded as 1, ’remaining’ class as 0.
As output of Data Preparation, three datasets were obtained: “EMPLOYEES_RECORDS” with 2077 rows and 26 features to use in the Data Description phase; “X_DATAFRAME” with 2034 rows and 40 features (12 numerical and 28 one-hot encoded) and ”Y_DATAFRAME” with only the target variable *Attrition*, both to use in the Data Prediction phase.

[12]As calculated in the salary benchmark phase

5.3 Data Description

Data Description is the activity of mining information from data using graphical tools. In this case, the purposes of this phase were two: obtaining a deep descriptive understanding of the dataset and having an early feedback on the features that could influence the target variable, that is employee attrition. These goals were achieved by plotting variables with respect to the two target classes, in order to detect different behaviors, and by performing a Pearson Correlation analysis with the target variable.

5.3.1 Results

Analyzing the target variable, in the dataset there are 237 observations labeled as 'leaving' and 1797 labeled as 'remaining'. Attritioner employees in the dataset are 11,6%. The dataset is indeed clearly unbalanced towards the 'remaining' class. There are some features which are particularly influenced by the different attrition classes: *Fixed Remuneration (RAL), Age, Seniority, Average Amount of a Bonus, Salary Gap*. The average seniority level is very different between the two classes: leaving employees have a mean seniority of 7.5 years, while remaining employees more than 20 years. Seniority appears to be a retaining factor. *Age* leads to the same conclusions, since leavers are on average 15 years younger than remaining employees. Also *Average Amount of a Bonus* and *Fixed Remuneration (RAL)* show the trend of providing lower averages for leaving employees. Moreover, considering *Salary Gap*, leaving employees show a negative gap of almost 3000 euros with respect to the market average. On the contrary, the remaining class has on average a positive gap.

The results of the correlation analysis are presented in Table 3.

Age and *Seniority* are the most correlated variables with the target one. Their correlation is negative, meaning that a lower age/seniority is more frequently associated with the value 1 in attrition, which is associated with the 'leaving' class.

Feature	Absolute Correlation	Sign
Age	0,397	-
Seniority	0,370	-
Average Time - Job Category Change	0,327	-
Number of Job Movements	0,218	-
Salary Gap	0,214	-
Number of Job Category Changes	0,138	+
Education_Inferior Diploma	0,125	-
Average Time - Job Bonus Award	0,117	-
Qualification_Manager	0,106	+

Table 3: Pearson correlation scores with the variable *Attrition*

5.3.2 Data Prediction

The Data Prediction phase is divided in the following sub-tasks: (a) Models Implementation, (b) Performance Metric Selection, (c) Training and Testing, (d) Performance Evaluation.

(a) Models Implementation phase is aimed at building the machine learning classification models to perform the pre-

diction task. Six models were chosen
to be implemented: Logistic Regression, Support Vector Machines, k-Nearest Neighbors, Classification Tree (CART), Random Forest, XGBoost. The first five algorithms were implemented from the *Sci-Kit Learn* package of Python 3.8, while the last one from the *xgboost* package. In order to implement the models, a pipeline was built which performed sequentially: Re-Sampling using a combination of SMOTE and under-sampling (since the dataset is unbalanced), Normalization with the function *StandardScaler* (only for models that re- quire it), Hyper-Parameters Tuning with the function *GridSearchCV*. Moreover, a features reduction was performed, where only the features most correlated with the target variable were kept. The implementation process is represented in Figure 3.
(b) Performance Metric Selection is the activity of choosing the evaluation metrics which

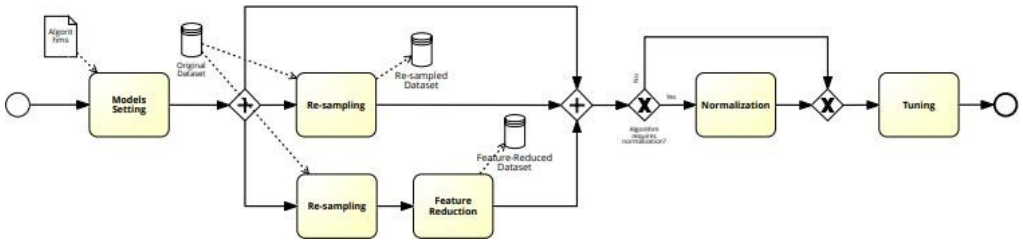


Figure 3: Models implementation process

can best illustrate models' performance. They are:

- Recall, since it detects the ability of a model in correctly identifying all the positives ('leaving' class);
- F1 Score and ROC-AUC, since they are not affected by class unbalancing.

(c) Training and Testing mean firstly to separate the data from which the model learns (training set), from the data used to assess models' performance. This split is performed in order to avoid overfitting[13] and in this case was realized with an 80:20 ratio[14].
(d) Performance Evaluation was performed on the chosen metrics for each model, both when trained on the original dataset, on the re-sampled dataset and on the features-reduced dataset, and then these results were compared each other.

5.4 Results

After re-sampling, size of the dataset was set to 1914, of which 1196 observations belonged to class 0 (62,5%) and 718 to class 1 (37,5%). The feature-reduced dataset instead had only the features of Table 3.
Characteristics of the three training and test sets are presented in Table 4. After performance

[13]Overfitting is the problem of a model which consists in exactly following the training data without having recognized what is a general trend and what is instead rumor
[14]80% training set, 20% test set

Dataset	Traning set size	Test set size	Ratio between 1 and 0	N° of Features
Original	1627	407	≈1:8	40
Re-sampled	1531	383	≈1:1,7	40
Feature-reduced	1531	383	≈1:1,7	9

Table 4: Characteristics of training and test sets

evaluation, the following conclusions were drawn:

- Tree-based models are the best from the point of view of each metric;
- Among tree-models, performance scores are really close when trained on the re-sampled dataset and on the feature-reduced one, whilst on the original data CART clearly outperforms the others. On the contrary, CART is always the best in processing times, since it is a simpler model;
- There is a strong improvement in performance for all models when re-sampling is implemented;
- Even better performance is reached by each model on the feature-reduced dataset, both in metrics scores and in processing times (except for Random Forest).

5.5 Final Model Selection

As a last step of the project, the best performing model was selected and its results were presented in greater detail.

5.5.1 Results

The best candidate model was the implementation of the Classification Tree CART on the featured-reduced dataset. This model achieved the following scores:

- Recall=1;
- F1 Score=0,98;
- ROC-AUC=0,994;
- Processing Time=10s.

Features importance for the best model is presented in Figure 4, where it can be seen that *Salary Gap* is the most important feature, followed by *Seniority*.

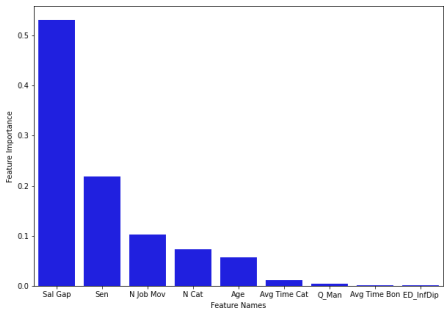


Figure 4: Features Importance of CART model trained on the feature-reduced dataset

6 Conclusions

This thesis dealt with the employee attrition problem, a huge concern for the majority of companies nowadays, and it represents an application of machine learning classification techniques on an international energy company employee dataset. The proposed framework is composed by 3 major phases which lead to the selection of the best performing prediction model on the dataset used. In this work, Classification Tree (CART) results as the best one, both considering predictive performance and processing times. Most relevant features are extracted and presented, which turn out to be *Salary Gap* and *Seniority*. Therefore, the introduction of the benchmark phase seems to be crucial since the feature *Salary Gap* was calculated based on the average salary values in output to the salary benchmark.
This framework could also be adopted by IT departments, which can automate the process in order to have always updated analytics on employees and employee attrition. Such information can indeed lead to great improvements if properly used by HR managers. As a matter of fact, only combining solid knowledge of HR processes and automated tools such as the one presented in this thesis generates the desired outcome of reducing churn rate of employees and hence attrition-related costs.
Given this scenario, two future developments appear to be of high potential:

- Going beyond predictive analytics in order to create a prescriptive framework which answers the question 'What can be done to retain this employee?' once an employee has been recognized as a potential leaver;
- Keeping the focus on the predictive field, studying the phenomenon with a vision not industry/company-specific to draw general conclusions about employee attrition prediction.

APPENDICES

A Methodology Flow

This appendix presents the methodology flow which was followed in order to carry out project's activities, as shown in Figure 5.

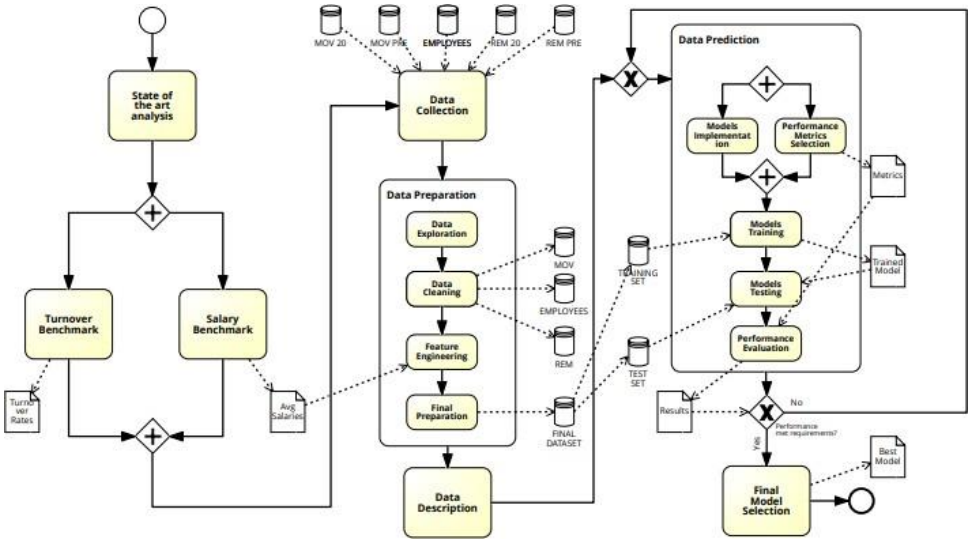


Figure 5: Methodology implemented in the project

B Final Model Selection

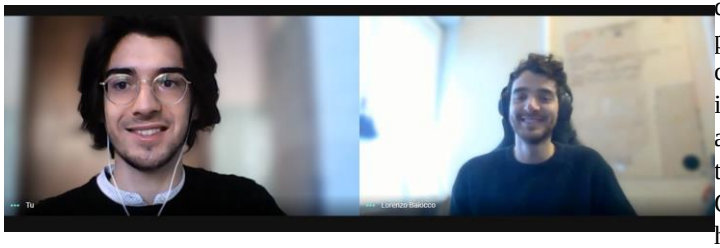
Final model selection was performed considering both the performance obtained on the selected metrics and the processing times. In Table 5, all the results are shown and the best model is highlighted.

Dataset	Model	Recall	F1-Score	ROC-AUC	Processing Times (s)
Original	Log. Regr.	0,61	0,64	0,785	11,7
	kNN	0,50	0,52	0,722	72
	SVM	0,72	0,71	0,839	390
	CART	1	0,86	0,979	23,8
	Rand. Forest	0,74	0,76	0,857	594
	XGBoost	0,87	0,81	0,917	240
Re-Sampled	Log. Regr.	0,96	0,91	0,937	10,3
	kNN	0,89	0,85	0,883	34,3
	SVM	0,95	0,91	0,934	384
	CART	0,99	0,96	0,972	12,4
	Rand. Forest	0,98	0,96	0,969	282
	XGBoost	0,97	0,95	0,965	246
Feature-Reduced	Log. Regr.	0,94	0,91	0,934	3,4
	kNN	0,94	0,91	0,935	18,5
	SVM	1	0,94	0,963	60
	CART	1	0,98	0,994	10
	Rand. Forest	1	0,98	0,990	366
	XGBoost	1	0,98	0,989	186

Table 5: Models' performance

C La mia esperienza in ELIS Consulting&Labs

Quest'anno, il contesto della pandemia mondiale da Covid-19 ha sicuramente stravolto e rivoluzionato qualsiasi esperienza. Il programma Junior Consulting non ne è stato esente ed ha dovuto reinventarsi in una nuova modalità interamente online. Non posso nascondere che la mia voglia di fare un'esperienza in prima persona di un ambiente di lavoro consulenziale come quello



di ELIS Consulting&Labs, di poter vivere l'open space e conoscere tanti colleghi da cui imparare e con cui magari fare anche amicizia fosse davvero tanta.
Ciò non è stato possibile, ma mi ha portato ad allenare una virtù che nella vita, ed in particolare in un momento delicato come l'attuale, è fondamentale: la resilienza. Se infatti da un lato non ho potuto avere un interfacciamento diretto con colleghi e clienti, dall'altra ho avuto la possibilità di fare esperienza dello smart working in ambito consulenziale: ho imparato a fare *team building* digitale, ho affinato l'abilità di presentazione online grazie alle frequenti interazioni col cliente, ho imparato a gestire orari e scadenze in autonomia. Avendo portato avanti tutte le attività di progetto, ho poi sviluppato delle competenze di project management in ambito Agile. Inoltre, il progetto mi ha portato ad acquisire un bagaglio di nuove competenze tecniche legate al mondo della *data analysis*, portandomi a riscoprire una forte passione verso la scienza statistica che supporta queste tecniche.
Il progetto è stato realizzato in squadra con Antonio (Responsabile di Progetto, Management Consultant presso ELIS) e Lorenzo (Team Leader, Data Scientist presso ELIS). Ringrazio i miei compagni di squadra non solo per avermi guidato nel progetto e nel mondo della *data analysis* grazie alla loro esperienza, ma anche per avermi da subito reso partecipe delle scelte progettuali ed aver dato valore al mio operato.

Fase del progetto	Competenze tecniche acquisite	Strumenti
State of the Art Analysis	Overview sulle tecniche di classificazione in ambito di Machine Learning. Conoscenza dello stato dell'arte riguardo al problema dell'*employee attrition prediction*	Scopus
Benchmark Analysis	Raccolta di dati, organizzazione ed estrazione di informazioni da essi attraverso metodologie grafiche. Conoscenza approfondita delle dinamiche del mercato salariale italiano	Excel pivot tables
Predictive Analysis	Fasi della pipeline di data analysis. Implementazione di modelli predittivi di classificazione	Jupyter Lab; Python, librerie pandas, numpy, seaborn, scikit-learn

Table 6: Competenze tecniche e strumenti acquisiti nelle fasi progettuali