Predicting Why Employees Leave

Andreas Rubin-Schwarz

***Abstract*—Measuring employee satisfaction is a tough and highly complex task. There are a lot of different dimensions in play and turning them into quantifiable format, less to say machine-readable information, can pose a challenge. Information from employee reviews, demographics, balanced scorecards and key performance indicators can offer a first gateway to understanding employee turnover. Modern machine learning algorithms offer tools to leverage this pool of information and extract actionable insights regarding employee behaviour. In this research a model is introduced using supervised and unsupervised algorithms to predict employee turnover. The novelty of the work is in the hybrid approach leveraging both unsupervised and supervised learning results. Based on recent research it was possible to increase the prediction outcome and to introduce a hybrid model that could work as a blueprint for similar tasks in this domain.**

***Keywords*—employee turnover, prediction, clustering, human resources**

I. Definition

A. Project Overview

According to the Bureau of Labor Statistics the median number of years that wage and salary workers had been with their current employer is 4.2 years [11]. While this number varies from industry to industry the story of an employee who sticks with one company for the entirety of a working life seems to be rather antiquated. This observation is combined with economical aspects. Employee turnover has been identified as a key issue for organizations because of its adverse impact on work place productivity and long term growth strategies [8]. One of the key issues with a high employee turnover rate combined with, but reaching beyond, cultural and sociological effects is the cost associated with it. Research shows that the replacement cost for an hourly worker can be as high as 50% of her annual salary. This number increases with the skill set of the worker up to 200% for senior-level workers and surges up to 400% for executive level positions [12]. It becomes obvious that the trend of shorter tenure in addition to high employee turnover rates can be a costly endeavour. Therefore it becomes increasingly important to acquire the necessary tools for employers to understand where its workforce is standing. Additional insights from employee reports, scorecards as well as general statistical information can offer prediction values for companies when it comes to the longevity of jobs. This research aims to predict the likelihood of an employee quitting her job based on available information. It tries to deliver a hybrid machine learning method to gain actionable insights on how to prevent a high employee turnover. The novel contribution of this paper is the usage of a two-layered approach using unsupervised and supervised machine learning algorithms to achieve a higher predictive outcome.

B. Literature Review

Employee turnover is a high impact topic and has been researched in both, management as well as psychology studies. Employee turnover can be defined as ”the gross movement of workers in and out of employment with respect to a given company” [6]. According to the Bureau of Labor Statistic there are two categories of turnover. Employee turnover is measured in terms of persons that quit their job, also known as voluntary turnover, and total separations. The subtraction of both numbers resulting in the amount for involuntary turnover. This research focuses on voluntary turnover and the reason behind it. Research suggests four categories of factors when it comes to employee turnover. Organization-wide factors, immediate work environment factors, job-related factors, and personal factors [5]. [4] indicate in their model a multi-layered approach to turnover motivators including traditional features such as job satisfaction, meeting of expectations or job involvement as well as newer attitudes like stress, psychological uncertainty, challenge, hindrance stressors or organizational context like company size, group cohesion and demography. Adding to the complexity of the field, research has shown, that turnover rates are also affected by moderating factors. General job availability, movement of capital and job satisfaction can interact with each other simultaneously to affect turnover [10]. Personal traits tend to have moderating effects as well. The relationship between turnover intentions and turnover for example can be moderated by various personality traits. This relationship was found to be stronger for employees with low self-monitoring, low risk aversion, and an internal locus of control [1]. Some research explored the relationship between job satisfaction and turnover and found it to be significant and consistent, but not particularly strong [7]. Capturing all these dimensions stays a problem. A more complete understanding of the psychology of the withdrawal decision process requires investigation beyond the replication of the satisfactionturnover relationship. [2] recommend a thorough comparison of several models and data exploration. Explicitly stating the need to continue model testing rather than simply correlating variables with turnover. [4] have found that most studies of voluntary turnover have one or two independent variables with voluntary turnover as the dependent variable. Machine learning algorithms are offering the opportunity to handle complex data structures and a large amount of variables. After capturing the needed data, machine learning can facilitate the ground truth hidden in massive data sets.

C. Problem Statement

Machine learning differentiates between two classification problems. In some cases the data set contains unstructured data and machine learning is used to find clusters or classes. In other cases there are certain already predefined classes or labels.

The task is to search for relationships or rules to distinguish one class from another. The first category of classification is called unsupervised learning. The second category falls into supervised learning. This research deals with a binary prediction task, to detect whether an employee has left the company or is still employed. But since employee turnover research is maneuvering in a highly interdependent context, the underlying relationship between features are important. In an attempt to create employee clusters and to use them in the prediction task, this research wants to offer a way of visualizing and categorizing employee behaviour. The problem to be solved is detecting the key elements of employee tenure and predicting whether an employee might be quitting her job. The setup of this research can be seen as a classification problem. Based on a set of features our solution should be able to determine if an employee quits or stays with a company. The core path to solving this problem is a supervised learning approach which will test the relationship between our independent variables and our dependent variable (did an employee leave or stay). The high-level tasks involved are the following:

1) Use statistical methods such as descriptive analysis, regression and correlation to lay groundwork.
2) Train a clustering algorithm to group employees into different segments.
3) Train a classifier that determines whether an employee has left.
4) Compare results against a benchmark model.

As to the supervised learning, this research will incorporate a Linear Support Vector Classifier, Logistic Regression model, Decision Tree Classifier and more sophisticated algorithms such as a Support Vector Classifier with polynomial or RBF kernel as well as ensemble methods like Random Forest Classifier. For unsupervised learning K-Means Clustering and Gaussian Mixture Model are on the shortlist.

A desirable solution can be quantified by correct detection of potential job quitters and the amount of created intervention opportunities. This problem affects companies of all sizes. Although it can be said that based on statistics there are certain industry traits that have a higher tenure the tendency to switch jobs has increased over time. The boundaries of this problem are the underlying mechanisms of every individual company. It is also important to mention that prediction is not equal to prevention. Company culture, direct reports and individual needs might turn out to be far more complicated than a list of features. However, if this problem is detected and fixed the path to a more successful and sustainable company culture is given. It shall be stated that a mechanism such as this can only work as a supplement to human interaction and empathy skills.

D. Solution Statement

First an in-depth exploratory analysis is performed in an attempt to explain the underlying relationships in the data set. Afterwards the problem is approached on two dimensions by combining unsupervised and supervised learning algorithms:

1) Two clustering algorithms namely K-Means or Gaussian Mixture Model will be discussed in an attempt to add more semantic to the employee set. Since the problem is based in a sociological domain, the more we can derive from our feature set, the more human interaction and perception we can gather, the better. By clustering, we might be able to hand additional information to the supervised algorithm for a better prediction down the road.
2) Afterwards features and additional cluster-information are used to predict whether an employee left or stayed. Given that the data is labeled and considering the amount of data points, a range of algorithms is discussed starting with more basic solutions and working towards more complex methods.

The solution is trying to implement a model that will empower employers not only to detect, which of their own employees is about to leave the company but also determine which of the features contributes the most to a happy company culture. The success of the final solution however will be measured by its predicting accuracy on a held-out test set.

E. Benchmark

As mentioned earlier, a lot of research and work has been performed to retain talent and decrease employee turnover rates. From corporate funded research and analytics services of companies [12] to management study classics [2] or more recent research [8]. There are several methods for predicting employee turnover rate and results that can be taken as benchmark models. In this research the machine learning approach of [8] serves as a benchmark. Their research performed predictive tasks on company information with an area under the curve (AUC) score of .86 on hold-out data implementing an Extreme Gradient Boosting model.

1) Evaluation Metric: Area under the curve (AUC) is a common evaluation metric for binary classification problems. Its value is between 0 to 1 and describes the accuracy of a binary classification based on its true positive values. An AUC score is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. [1] To put this more into perspective AUC can be seen as a plot of the true positive rate vs. the false positive rate where the threshold value for classifying an item as 0 or is increased from 0 to 1.[2] If the classifier is very good, the true positive rate will increase quickly and the area under the curve will be close to 1. If the classifier is no better than random guessing, the true positive rate will increase linearly with the false positive rate and the area under the curve will be around 0.5 [3].

2) AUC vs. F1: There is some controversy about the value of AUC scores. Especially when it comes to precision and recall it is possible for a classifier to have a low recall but a very high AUC score. To avoid this, F1 scoring can be used which is a way to incorporate both precision and recall. For a F1 score to be high, both, precision and recall have to be high. The problem of a low recall score will be addressed for the leading model when testing for robustness.

[1] https://en.wikipedia.org/wiki/Receiver_operating_characteristic

[2] https://www.kaggle.com/wiki/AreaUnderCurve

II. Analysis

In a data set published on Kaggle, information on current and former employees is offered plus key features of their employment status. Unfortunately there is no available code book besides a brief information on the available inputs. The data set contains 14,999 data points and 10 variables.

Since there is no additional information regarding the data set we are starting with some assumptions about the variables based on their description. Information that's comprised in this data set seems to include a measure of employee satisfaction level, evaluation scores, amount of projects, average monthly hours spend on the job, overall time spent at the company, whether the employee had a work accident or received a promotion within the last 5 years, a description of the department the employee works for and a salary indicator. In addition there is a variable called left that indicates, whether an employee is still actively working for this company or not. This indicator will be seen as the dependent or target variable that is being predicted.

A. Data Exploration

As mentioned earlier, there is no code-book that can be linked to this data set. Since there is no additional information regarding the data set some assumptions about the variables are made based on their description. A sample of the data looks like this:

TABLE I. Sample of Data Set (n = 5)

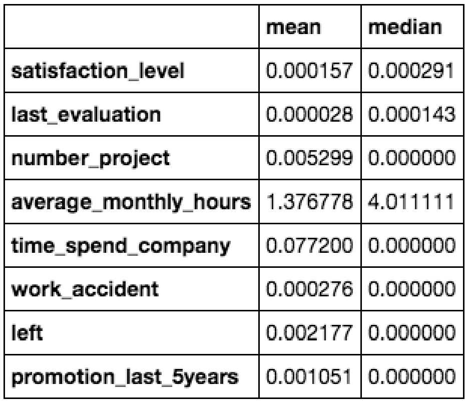
	sat_level	evaluation	projects	avg_hours	time	accident	left	promotion	dept	salary
0	0.3800	0.5300	2	157	3	0	1	0	sales	low
1	0.8000	0.8600	5	262	6	0	1	0	sales	medium
2	0.1100	0.8800	7	272	4	0	1	0	sales	medium
3	0.7200	0.8700	5	223	5	0	1	0	sales	low
4	0.3700	0.5200	2	159	3	0	1	0	sales	low

A brief summary of the numeric values in the data set offers following insights about the variables:

- **satisfaction_level**: Most likely indicating a high satisfaction level. [continuous value in the range of 0 and 1]
- **last_evaluation**: Most likely the last evaluation score of given employee. [continuous numeric value between 0 and 1]
- **number_project**: The amount of projects an employee has. [discrete numeric value between 2 and 7]
- **average_monthly_hours**: The hours an employee works per month. [continous numeric value between 96 and 310]
- **time_spend_company**: Given current working statistics we are most likely dealing with years. [discrete numeric value between 2 and 10]
- **work_accident**: Numeric expression of boolean value whether an employee had a work accident. [discrete: 1 = true, 0 = false]
- **promotion_last_5years**: Numeric expression of boolean value whether an employee was promoted within the last 5 years. [discrete: 1 = true, 0 = false]
- **department**: Categorical variable describing the position of an employee. [discrete: sales, accounting, hr, technical, support, management, IT, product_mng, marketing, RanD]
- **salary**: Categorical variable indicating salary level of employee. [discrete: low, medium, high]
- **left**: Label if person left or not. [discrete: 1 = true, 0 false]

1) Mean vs median: In order to decide whether to pick the mean or median values for comparison, the variance of each option is computed.

TABLE II. Variance of Mean vs Median



	mean	median
satisfaction_level	0.000157	0.000291
last_evaluation	0.000028	0.000143
number_project	0.005299	0.000000
average_monthly_hours	1.376778	4.011111
time_spend_company	0.077200	0.000000
work_accident	0.000276	0.000000
left	0.002177	0.000000
promotion_last_5years	0.001051	0.000000

Tab. II shows the result of variance for average and median values of each numeric variable. There is clearly more information gain from the mean values in our data set. Therefore average values will be used for our analysis where needed. A summary computation of standard deviation, minimum and maximum values as well as the numeric quartile ranges indicates that the data set is already quite balanced. Based on the numeric information about mean, minimum and maximum values it can already be said that the numeric values seem to contain only little anomalies when it comes to tendency.

2) Univariate Analysis: To substantiate these findings all variables are explored one-by-one. The method used depends on whether the variable is continuous or discrete. For continuous variables the spread and central tendency is important. For discrete variables frequency and histograms will be used to gain additional information. The standard deviation values indicate a balanced dispersion (see Tab. III). In order to visualize the findings density plots[3] are created to investigate the distribution. Following observations can be made:

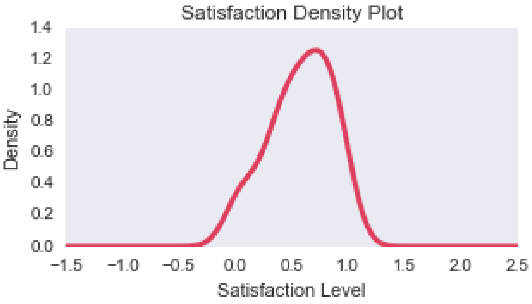
With some smoothing, the satisfaction_level seems to be pretty normal in regards of distribution (Fig. I). Both

[3] The density curve describes the relative likelihood for a random variable to take on a given value.

TABLE III. STANDARD DEVIATION, MINIMUM AND MAXIMUM VALUES OF KEY VARIABLES

feature	std	min	max
satisfaction_level	0.25	0.09	1.0
last_evaluation	0.17	0.36	1.0
number_project	1.23	2.00	7.0
average_monthly_hours	49.94	96.00	310.0
time_spend_company	1.46	2.00	10.0
work_accident	0.35	0.00	1.0
left	0.43	0.00	1.0
promotion_last_5years	0.14	0.00	1.0

FIG. I. DENSITY PLOT OF SATISFACTION LEVEL



last_evaluation (Fig. II) as well as average_monthly_hours (Fig. III) have two peaks. Yet there seems to be no anomaly in the distribution.

FIG. IV. FREQUENCY PLOT OF NUMBER OF PROJECTS

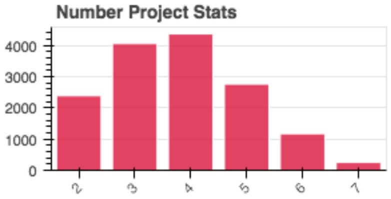


FIG. II. DENSITY PLOT OF LAST EVALUATION SCORE

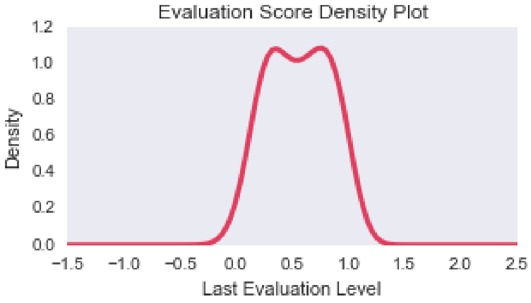


FIG. III. DENSITY PLOT OF MONTHLY WORKING HOURS

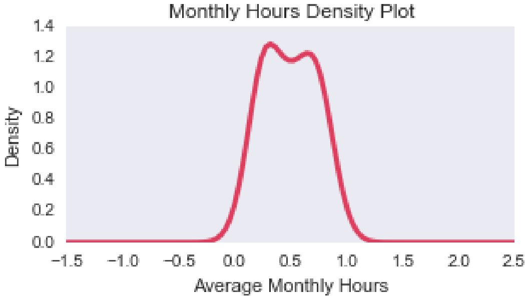
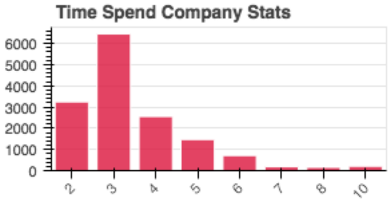
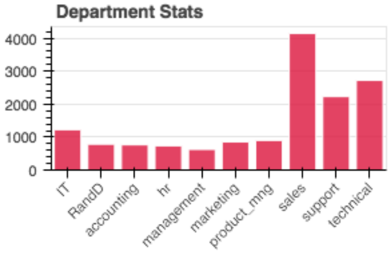


FIG. V. FREQUENCY PLOT OF TIME SPEND AT COMPANY



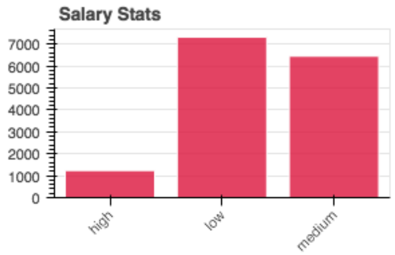
The discrete variables are explored with frequency plots. Due to the amount of only seven discrete variables it is possible to plot all variables and by doing so gain first insights about the data. Looking at the frequency plots the number_project (Fig. IV) and time_spend_company (Fig. V) distribution have a slight positive skew. It can also be seen that the most employees tend to have three or four projects and tenure seems to level around three years.

FIG. VI. FREQUENCY COUNT OF EMPLOYEES IN DEPARTMENTS



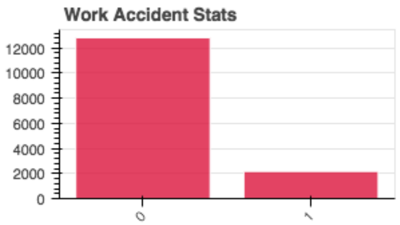
When it comes to department, the most employees are employed in sales with technical coming second and support being third on the list. It is interesting that technical and IT are separate divisions (Fig. VI).

FIG. VII. FREQUENCY PLOT OF SALARY LEVEL



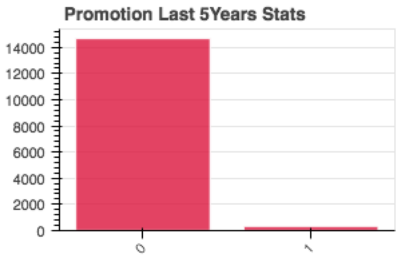
The salary variable shows a clear skew towards low and medium wages which was to be expected. The majority of the workforce described in this data set has a medium to low salary level (Fig. VII).

FIG. VIII. FREQUENCY PLOT OF WORK-RELATED ACCIDENTS



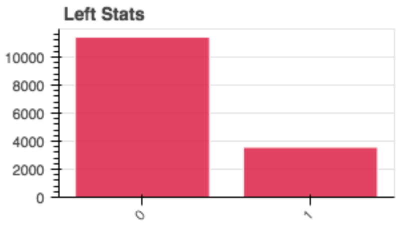
When it comes to work accidents, the data shows about fourteen in 100 employees has had a work related accident (Fig. VIII).

FIG. IX. FREQUENCY PLOT OF PROMOTION WITHIN THE LAST 5 YEARS



Even more severe is the skew when it comes to promotions. Only two out of 100 employees have had a promotion in the last five years (Fig. IX).

FIG. X. FREQUENCY PLOT OF EMPLOYEE TURNOVER



A slightly less severe but still significant skew is happening in the dependent variable. This needs to be taken into account when splitting the data set for training, cross-validation and testing (Fig. X).

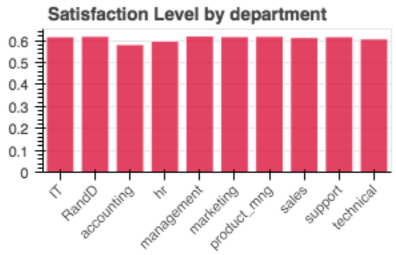
3) Data Context: To get a firmer grasp on some of the information hidden in the data set, it is helpful to ask questions regarding the context of the data. Simple questions help putting information into perspective and offer a gateway to understanding the data. Questions that are being answered by this section are:

- How many people left their company?
- What is the average tenure?
- How do variables change by department?
    - Which department works the most hours?
    - Where do most work accidents happen?
    - What is the satisfaction level by department?

An overall of 3,571 Persons left their company which is around 24%. It is important to note that there needs to be a similar distribution of this variable in the training and testing sets used later. It shows that people in our data set tend to leave the company earlier than the average mentioned in the first paragraph. This could indicate that the population we're looking at is either younger than the average work force or, more likely, that we're not representing state and federal employees which tend to have a higher average tenure ($>$ 8 years) [11].

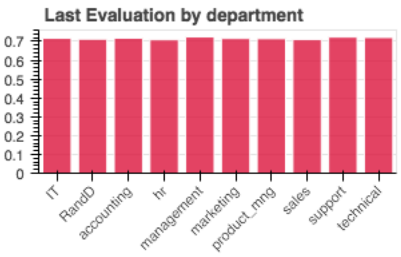
4) Exploration by department: In order to get a better grasp on the information that is contained in the data set, we'll be looking at average values of the predictors through the eyes of each department. The mean values per department are calculated for all discrete variables and plotted accordingly.

FIG. XI. AVERAGE OF SATISFACTION LEVEL BY DEPARTMENT



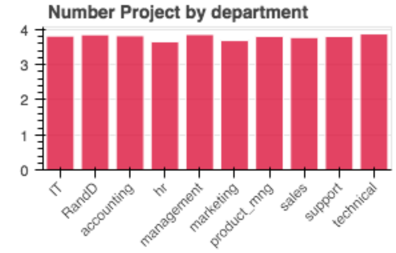
The satisfaction level seems to be quite consistent over all departments. Only accounting and HR have a slight below average satisfaction level.

FIG. XII. AVERAGE EVALUATION SCORE BY DEPARTMENT



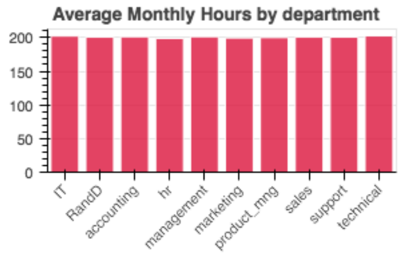
Evaluation levels are on a consistent level over all departments, which speaks for a well-balanced evaluation metric.

FIG. XIII. AVERAGE NUMBER OF PROJECTS BY DEPARTMENT



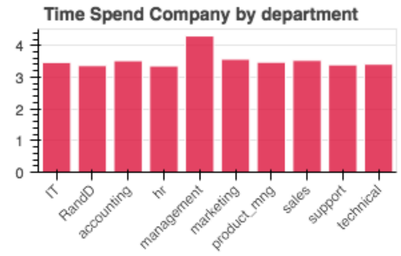
HR and marketing seem to have a slightly below average number of projects.

FIG. XIV. AVERAGE MONTHLY WORKING HOURS BY DEPARTMENT



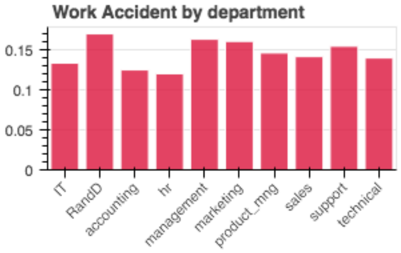
All departments seem to clock in the same amount of time on a monthly basis.

FIG. XV. AVERAGE TIME SPEND AT COMPANY BY DEPARTMENT



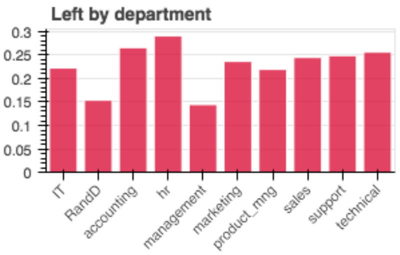
Management is the clear front runner, when it comes to tenure. This makes sense, since a management position comes with greater responsibility and lock-in effects.

FIG. XVI. AVERAGE AMOUNT OF WORK-RELATED ACCIDENTS BY DEPARTMENT



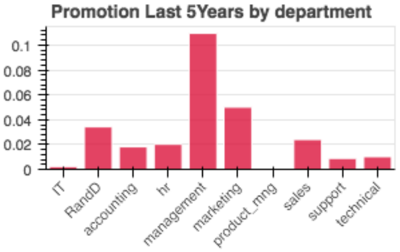
R&D, Management, Marketing and Support seem to have the most hazardous work environment. With accounting and HR being on the safety first side.

FIG. XVII. AVERAGE OF EMPLOYEE TURNOVER BY DEPARTMENT



Management and R&D have the highest average loyalty rate. Accounting and HR on the other hand are fluctuating quite a bit.

FIG. XVIII. AVERAGE OF PROMOTIONS IN THE LAST 5 YEARS BY DEPARTMENT



Management have by far the highest promotion rate. This might speak to the being of a management position. In order to get to a management position you need to have a promotion. In marketing almost one out of two has had a promotion within the last 5 years, in the R&D department over 30% has advanced on a professional level in the last 5 years. The support and technical departments are on the lowest spectrum.

5) Bi-Variate Analysis: It's good practice to get a grasp on how the variables are behaving in concert with each other. The relationship between two variables can tell us a lot about possible hypotheses that might be hidden in the data. In order to leverage this part of the analysis to its full extend One-Hot encoding will be performed for all discrete character filled variables, which is a step featured in data preprocessing.

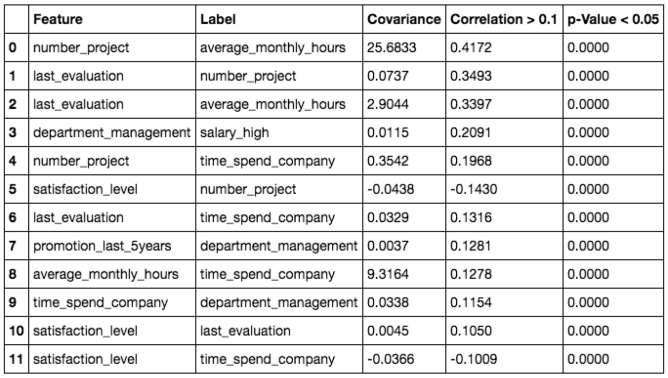
B. Regression & Correlation

After encoding and preprocessing the data set regression and correlation calculations were performed. This is an important step to understanding what measures might be helpful when trying to prevent a high leaving rate. Just based on the first correlation results there are already a couple of interesting observations:

1) A high satisfaction level seems to lead to a lower leaving rate.
2) Employees with higher salary are less inclined to leave.
3) On the other hand low salary employees seem to leave the company more often.
4) There's a positive correlation between the time a person spends at a company and the fact if they left.

In order to get more granular highly correlated (strength $> .1$) and highly significant (p-Value > 0.05) connections are being investigated.

TABLE IV. CORRELATION TABLE (STRENGTH $> .1$ & $p > .05$)



	Feature	Label	Covariance	Correlation > 0.1	p-Value < 0.05
0	number_project	average_monthly_hours	25.6833	0.4172	0.0000
1	last_evaluation	number_project	0.0737	0.3493	0.0000
2	last_evaluation	average_monthly_hours	2.9044	0.3397	0.0000
3	department_management	salary_high	0.0115	0.2091	0.0000
4	number_project	time_spend_company	0.3542	0.1968	0.0000
5	satisfaction_level	number_project	-0.0438	-0.1430	0.0000
6	last_evaluation	time_spend_company	0.0329	0.1316	0.0000
7	promotion_last_5years	department_management	0.0037	0.1281	0.0000
8	average_monthly_hours	time_spend_company	9.3164	0.1278	0.0000
9	time_spend_company	department_management	0.0338	0.1154	0.0000
10	satisfaction_level	last_evaluation	0.0045	0.1050	0.0000
11	satisfaction_level	time_spend_company	-0.0366	-0.1009	0.0000

Based on this correlation analysis we can infer following observations:

- Employees with a higher number of project tend to have a significant higher work-load per month.
- A high evaluation level leads to an increase in the amount of projects (and therefore an increased monthly work-load).
- People from the management department tend to have a higher salary.
- An increase of number of project tends to a higher tenure rate.
- A low number of projects has a negative correlation with the satisfaction level.
- If the last evaluation was higher, the person was more likely to stay longer with the company.
- As mentioned above, people from the management department tend to have a higher promotion rate.
- A high effort, measured by average monthly hours, leads to a significant higher retention rate.
- People in management jobs have a higher tenure.
- A high last evaluation has a positive impact on the satisfaction level.
- A low satisfaction level has a negative correlation with the time an employee stays at a company.

C. Outlier Detection

One problem that might occur in a data set are outliers. Outliers can lead to wrong estimations and poor performance on some of the algorithms. An outlier is a data point that seems to be out of the normal when it comes to a variable or even multiple variables at once. Usually outliers can be uni-variate or multivariate which means that their abnormal values can either happen on one dimension or multiple variable dimensions together. Going on, the focus will be on uni-variate outliers.

The initial value to detect outliers used in this analysis is a 1.5 multiple of the Inner Quartile Range. Based on this definition there are 1,282 outliers in only one variable (time_spend_company). The outliers amount for 8.55% of the overall data set. In order to determine if these outliers should be removed from the data set a broader range is used. By increasing the tolerance for outliers, additional variance might be added but it gives us the opportunity to keep more data for our algorithms. Just widening the step range by .5 reduces the outlier-rate by just shy of 5%. To keep as much data as possible an even closer look is taken. The minimum and maximum IQR boundaries that distinguish outliers per our definition is 3 and 4. Taking into account that this boundary is quite slim all possible outliers of this variable will be used for the analysis.

III. Algorithms and Techniques

The list of possible algorithms for this problem is long. There is a large set of possibilities on how to tackle this task but our key metrices to narrow down the list of algorithms will be data-size, computational efficiency and cross-validation scores on the training set. To come closer to a decision when it comes to the potential algorithm the prerequisites of our setup have to be taken into account. To apply any kind of machine learning technique a data set with over 50 data points is required. In this case with around 15,000 entries, that prerequisite is met. The data is labelled, which puts this research in the realm of supervised learning. Yet, given the complexity of the domain the research aims at engineering an additional variable through unsupervised learning. By doing so an additional discrete variable is created that allots employees based on their attributes to a certain cluster. This section touches base on the idea behind various supervised and unsupervised classification algorithms.

A. Supervised Algorithms

First supervised learning algorithms are explored, detailing some of their features and drawbacks. While some models work better on large sets, some of them are better suited for a small set of training data. A common threshold for machine learning algorithms is 100,000 data points. Since the amount of data for this problem lies way below this threshold, models that perform well on a small set size are preferred. it can be assumed that time wont be too much of an issue since all calculations happen offline.

1) Support Vector Machine: Support Vector Machines (SVM) are supervised learning algorithms implementing the principles of statistical learning theory. These will be used in a linear and non-linear implementation to account for a higher complexity of the data. SVM tend to have a high accuracy while maintaining a fairly low variance, hence are more unlikely to overfit. This model works really well in complicated domains where there is a clear margin separation. SVM construct a hyper-plane to distinguish the data points from each other. The intuition being that the larger the margin from the closest data points of each class the better the separation criteria. Even if the data turns out to not be linearly separable, this model usually works quite well. However, if the data contains a lot of noise or the amount gets too large, the model tends to perform poorly or very slow. Considering that in a running environment the data base tends to grow this approach might become unfeasible if implemented in a live system. In order to get the best result, key parameters that will be tested for linear SVM are the penalty parameter C of the error term, the used loss-function and the amount of maximum iterations to run as well as l2 penalty function. The implementation with a more complex kernel will also take into account a range of different kernel functions and a range of polynomial degrees for increased complexity.

2) Logistic Regression: Logistic Regression (LR) is one of the basic linear models for classification. It is best used to predict binary or discrete dependent variables. LR is usually a quick and easy solution for machine-learning problems. It's implementation is fairly similar to SVM with the great upside to work better on large data sets. Additional advantages being that there are a lot of ways to regularize the model and correlation of features doesn't matter as much. Another one of its main advantages is, that new data can easily be added and the model can be updated in an ongoing process. Its flexibility makes it a perfect choice for a running system, that should either be adjusted over time to model better or infused with more data. A downside of LR, is that it might have difficulties with binary features. In this data set that contains a substantial set of categorical features this might become a problem. For Logistic Regression tuning involves the penalty parameter C of the error term as well. In addition several solvers are being used as well as a variation on whether the algorithm is able to reuse the solution of the previous call to fit as initialization or if it has to erase the previous solution also known as warm start.

3) Decision Tree Classifier: Another viable option is a Decision Tree model (DT). Decision Trees are fairly easy to explain and interpret. They are also pretty robust against outliers and can be protected against overfitting through pruning. However, Decision Trees don't support online learning and have to be rebuild every time new information comes in. This might matter, if the used data is regularly updated and not only once a quarter. If an accurate turnover prediction model is needed, that works year round on recent data and updated models, Decision Trees might have a disadvantage. Decision Tree parameters that are taken into account are the function to calculate the splitting criterion measure, the number of features to consider when looking for the best split as well as the maximum depth a tree can have.

4) Random Forest: There are ensemble methods that incorporate trees such as Random Forest (RF). RF could prove itself with its ability to accept non-linear features. Other than LR it can handle categorical features very well. It's also well suited for high dimensional spaces in case more features are added or engineered and large numbers of training examples. As mentioned this can be important for future application, once the data base grows. A major downside of RF-models though is its lack of sensitivity towards correlated features. With correlated features, strong features might end up with low scores. Parameters that will be tweaked to find the best working solution are different criterions to measure splitting performance, the ability to bootstrap samples or not, several functions to calculate the maximum features per tree, a maximum depth of the tree, an array of minimum splits per sample and several options for the maximum of leaf nodes a tree can have. In addition, this algorithm's performance will be examined with or without the ability to perform a warm start as well.

Additional information on set, selection and performance will be shown in Chapter V.E.

B. Unsupervised Algorithms

One problem this research is facing when it comes to unsupervised algorithms is the mixture of discrete and continuous variables. Most algorithms mentioned here are better suited for continuous variables and might perform less accurate when working with categorical features. This segment therefore is meant to be a supplement to the supervised learning models. Its efficiency will be tested based on the cross-validation score before and after the feature employee_cluster has been added. Clustering the data has another real-life application. Categorizing employees will give the opportunity to create more tailored solutions for each employee cluster. Following are two algorithms that have been on the shortlist for this task.

1) Gaussian Mixture Model: Gaussian Mixture Models (GMM) are probabilistic models for representing normally distributed subpopulations within an overall population. GMM are used a lot when the underlying populations can be explained by a normal distribution and many heterogeneous populations are available. As an example based on [9], we can look at the average evaluation scoring for people in the different departments: R&D, Accounting, HR, Management, Marketing, Product Management, Sales, Support and Technical. It can be assumed the evaluation score distribution is slightly different within each department and it follows a normal distribution. The weighting factor could be the percentage of the population that comes from each department as defined above. This would be a 9-point Guassian Mixture Model. The key benefits of GMM are density estimations for each cluster, a certain flexibility when choosing the component distribution and the possibility of soft classification. GMM is a Bayesian approach to clustering. It introduces the ability of soft clustering, which means that data points can be part of more than one cluster. The algorithm also calculates the probability of the data point belonging to a certain center. It is also known to reflect real-world scenarios in a good way.

2) K-Means Clustering: In general K-Means can be seen as a special case of GMM in which each cluster's covariance along all dimensions approaches 0. Meaning each data point will be assigned to exactly one cluster. Some of its key advantages are that K-Means is robust and easy to understand. It is computational efficient and delivers a great result when data points are distinct or groups within are well separated from each other.

Because we're dealing with a real world scenario and a data set where it sometimes might be hard to distinguish between data points, we'll go ahead with a GMM implementation. This will also give us the opportunity to soft-label our data-points and refine the clustering at a later point.

C. Benchmark

As mentioned before, there are valid methods for predicting employee turnover rate. These results can be taken as benchmarks. In this research the machine learning approach of [8]. Their research performed predictive tasks on company information with AUC scoring of .86 on hold-out data as a best result.

1) Evaluation Metric: AUC scoring can be controversial. To account for this, F1 scoring will be implemented on the test set, once cluster-enhancement and GridSearch have been implemented to make sure the model works well with high precision and recall scores.

2) Evaluation Method: First the algorithms will perform a 5-fold cross-validation on 80% of the data. The average score will be the indicator for the best performers. Afterwards the whole data set will be clustered and the supervised learning algorithms are tested on their performance again through cross-validation. To avoid overfitting, which can be a problem especially in tree-based models, the same procedure is applied to the 20% hold out set. Both measures, AUC as well as F1 will be reported and evaluated.

IV. Methodology

A. Data Preprocessing

In this section all steps leading to a clean and usable data set are outlined. This includes: detection and treatment of missing values as well as outliers, encoding of categorical variables and if necessary normalization of numeric ones. Finally this section gives an overview on feature importance.

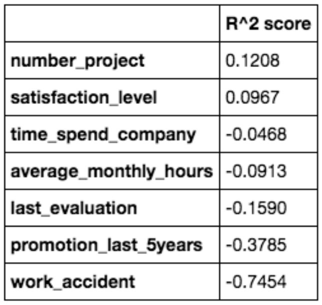
1) Missing values and outlier detection: The data doesn't contain any missing values. In an attempt to trim the data into a more concise state, outliers with a value greater or smaller than two times the Inner Quartile range have been detected. Yet, since the boundaries were quite slim all outliers are included in the prediction process. The core steps and thoughts taken in this segment are reflected in Section II.C.

2) One-Hot Encoding: Since some algorithms require all variables to be numeric, all discrete character-based variables were encoded and so-called dummy variables have been created. This gives all used supervised learning algorithms the chance to read in all information that can be derived from these categories.

3) Feature normalization: In order to get all data on the same scale, min-max normalization is performed on all numeric variables. Ensuring standardised feature values implicitly weight all features equally in their representation. Training runs in this research have shown that especially algorithms such as Logistic Regression and LinearSVC perform better with normalized data.

4) Feature Relevance: One interesting thought to consider is, if one of the categories is relevant for understanding why an employee stays with a company. It might be possible that a category can be predicted by the other categories and therefore contains only limited information regarding the final goal. In simple terms, it might be possible to for example predict the number of projects based on the constellation of all other categories. This determination can be made quite easily by training a supervised regression learner on a subset of the data with one feature removed, and then score how well that model can predict the removed feature. The coefficient of determination, R^2, is being used to determine the regression score. R^2 has a positive score range between 0 and 1, with 1 being a perfect fit. A negative R^2 implies the model fails to fit the data. In this step One-Hot encoded features are not predicted since its value can easily be predicted by looking at the rest of the encoded variables. To gain better insights in the relevance of each category and to understand if an employee will stay at a company, each category is selected and dropped from the data set. The remaining variables are being used to predict the dropped variable. If prediction accuracy is high for the variable, its relevance for the prediction process might be low. Tab. V shows the resulting R^2 score table.

TABLE V. FEATURE RELEVANCE R^2 SCORE WITH DECISIONTREEREGRESSOR



	R^2 score
number_project	0.1208
satisfaction_level	0.0967
time_spend_company	-0.0468
average_monthly_hours	-0.0913
last_evaluation	-0.1590
promotion_last_5years	-0.3785
work_accident	-0.7454

The Decision Tree Regressor failed to predict all categories but number_of_project and satisfaction_level. Which means all these categories are important. The R^2 scores for number_of_project and satisfaction_level (both below .2) show that the regression model had difficulties predicting the values as well. We can therefore infer that all numeric categories contain valuable information and should be used.

V. IMPLEMENTATION

After making all features machine readable, a model needs to be selected. Because the task at hand is to find out whether an employee left the company or is still employed, this problem is approached primarily as a classification task. Since the amount of labeled data is in the low-mid-range with below 100,000 data points, the prediction task will be started with a Linear Support Vector Machine. There are more sophisticated machine learning algorithm but it's always a good idea to start at the bottom and get more complex as research continues. The second model is a Decision Tree Classifier, because of its simplicity and computational efficiency. As a third contender a more sophisticated Support Vector Machine is implemented that leverages additional dimensions through an RBF[4] kernel. Finally a Random Forest Classifier will be used to predict. Since sklearn makes it quite easy to add additional estimators training and cross-validation is implemented once and the process adapted in a loop for all models.

A. Setup

The prediction process is as follows:

1) The data set with all machine readable information is split into a training and a testing set. 20% of the original datapoints are left out for testing purposes to validate the final model. In this step the stratify option is being used in order to keep a similar distribution of label data in testing and training set.
2) The training features set will be split into independent (predictors) and dependent (label) variables. The predictors will be used to predict the label.
3) A random state will be set which will help with the reproducibility of the prediction process.
4) In order to avoid overfitting k-fold cross-validation with k = 5 is performed. This means the training set is split in five separate buckets. Four of these buckets are used to train and the left out bucket is used for testing. CV performs this k-times. The average score will give a better indicator on how the model will perform on the testing set or in a real-life application.
5) **Base Prediction:** Prediction takes place with basic settings.
6) **Clustering:** An unsupervised learning algorithm is implemented to classify employees based on their variables and engineer an additional discrete feature for prediction (and pro-active intervention) purposes.
7) **Cluster Prediction:** Prediction takes place with additional cluster information.
8) In order to find the best possible setup for the supervised learning algorithms grid search is performed on all algorithms.
9) **Fine Tuned Prediction:** The fine tuned parameter setup is being used to predict on the cluster enhanced data set.
10) **Testing:** All trained models are being used on the 20% hold-out data. The highest prediction score in this process is being recorded.

[4] RBF stands for radial basis function. RBF kernel is a popular kernel function for Support Vector Machine classification and sometimes used as a similarity measure.

B. Base Prediction

The first prediction run takes place with default settings. For scoring, AUC is measured over a 5-fold cross-validation on the training set. Tab. VI shows the score for all used algorithms.

TABLE VI. PREDICTION WITH NO ENHANCEMENT OR TUNING

Classifier	AUC	Time
RandomForestClassifier	**0.9904**	0.6606
DecisionTreeClassifier	0.9765	0.3931
SVC	0.9091	17.6686
LogisticRegression	0.8179	0.3413
LinearSVC	0.8173	0.6458

The Random Forest ensemble model has the best average cross validation AUC score with .9904. Fine-tuning of the algorithm parameters will happen later, once the unsupervised additions have been computed.

C. Clustering

Next an additional layer of context will be added through clustering employees into groups. By doing so a new discrete variable is created that might give additional insights into the data. For clustering this research chose GMM because of its features to allow soft classifications as performing well in real-world scenarios well. Because we're dealing with a real world scenario and a data set where it sometimes might be hard to distinguish between data points, we'll go ahead with a GMM implementation. In a running environment the clustering should be done offline to safe computation expenses. As mentioned earlier, combining discrete and continuous variables for this task is tricky. Therefore the focus of this supplemental method is on the continuous variables. Since the data is already normalized the next step is feature transformation in order to determine dimensions that are hidden in the data set.

1) Feature Transformation: As a first step towars dclustering employee data Principal Component Analysis (PCA) is performed on the data set to discover which dimensions about the data best maximize the variance. In addition to finding these dimensions, PCA will also report the explained variance ratio of each dimension. This suggests how much variance within the data is explained by that dimension alone and gives a good indicator about distinguishing performance of each given dimension. A component, also called dimension, from PCA can be considered a new feature of the space. However it is a composition of the original features present in the data. As shown in Table VII, the first 3 dimensions explain 78.45% of the variance. Adding dimension 4 pushes that value to just shy of 90%. It's interesting to see that the dimension that explains the most variance has mainly negative values and indicates a negative relationship between satisfaction_level and the other variables. Dimension 2 builds up on that fact, seeing a negative mix of satisfaction_level and last_evaluation score in a dependency situation with time_spend_company and number_project (see Fig. XIX).

FIG. XIX. OVERVIEW OF PRINCIPAL COMPONENT ANALYSIS WITH 5 DIMENSIONS

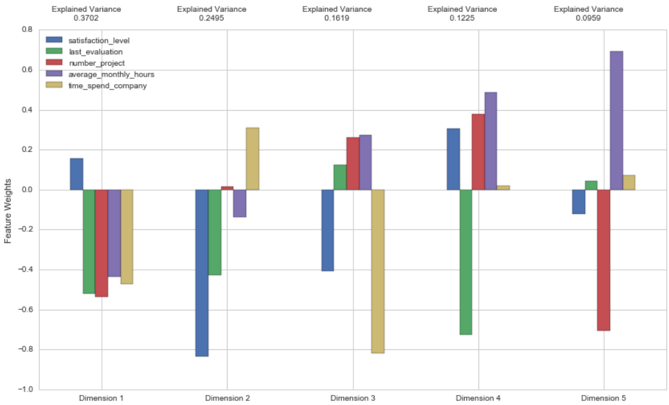


TABLE VII. CUMULATIVE EXPLAINED VARIANCE BY DIMENSION

	Dim1	Dim2	Dim3	Dim4	Dim5
Variance	0.3745	0.6439	0.7845	0.8973	1.0000

2) Silhouette Score: These insights are taken further by calculating the average silhouette score for each cluster setup. Silhouette analysis shows how well clusters separate and distinguish points. The measure has a range of [-1, 1] where +1 indicates that the separation of points is very well done and -1 indicates that the points might have been mislabeled. The silhouette analysis results can be seen in Tab. VIII.

The average silhouette score for all points shows that four components has the best separation power. Yet, the value is in the low positive range which might indicate that the company should add additional continuous features to describe its clusters better. Using the silhouette score as an indicator, employees will be separated into four classes.

3) Visualization of clusters: In order to get a better understanding on how this separation looks like, a scatter plot is created that marks both, clusters and its members. It is obvious that the separation is heavily influenced by the first two dimensions. A scatter plot of the first two dimensions gives a better understanding on how the data is structured (see Fig. XX). Since the silhouette score suggests four dimensions to use when clustering, the visual reflection of this plot is limited. In an attempt to shed more light on the cluster distribution a 3D plot is created (Fig. XXI) incorporating the third dimension values as well. Yet, because this research is using four dimensions to cluster the data it is hard to show the separation panes for all dimensions. The difficulty to properly distinguish data points however is a fact that is also reflected

TABLE VIII. SILHOUETTE SCORE WITH SEVERAL AMOUNTS OF COMPONENTS

	5 comps	4 comps	3 comps	2 comps
Score	0.2163	0.2221	0.2020	0.2091

FIG. XX. Cluster Representation in Two Dimensional Space.

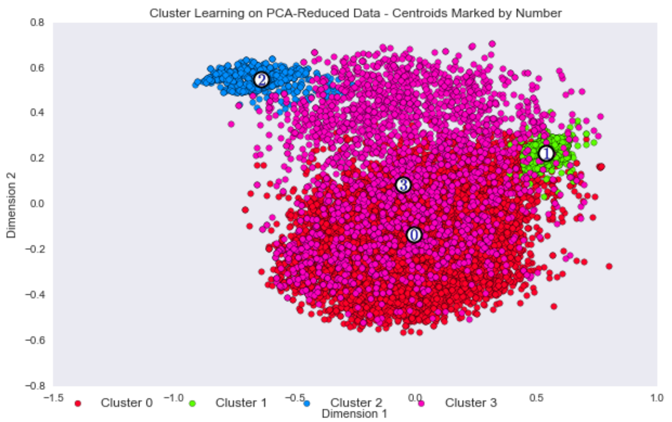
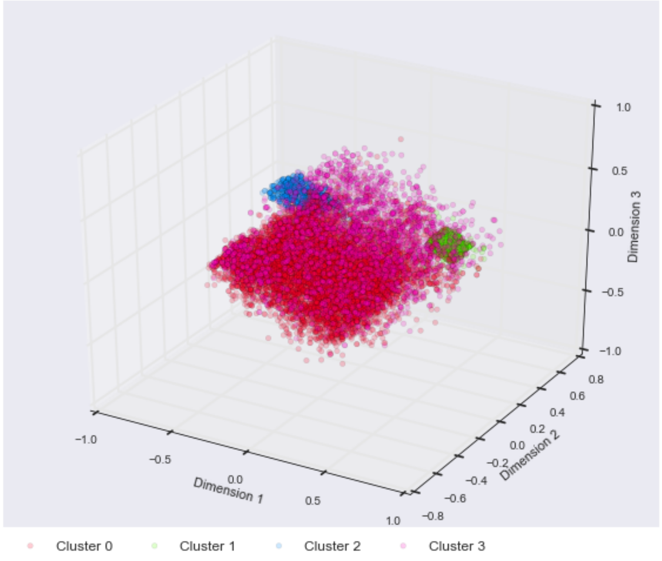


FIG. XXI. Cluster Representation in Three Dimensional Space.



by the .2221 silhouette score. With a higher silhouette score the clusters would be easier to separate.

D. Prediction on cluster-enhanced data

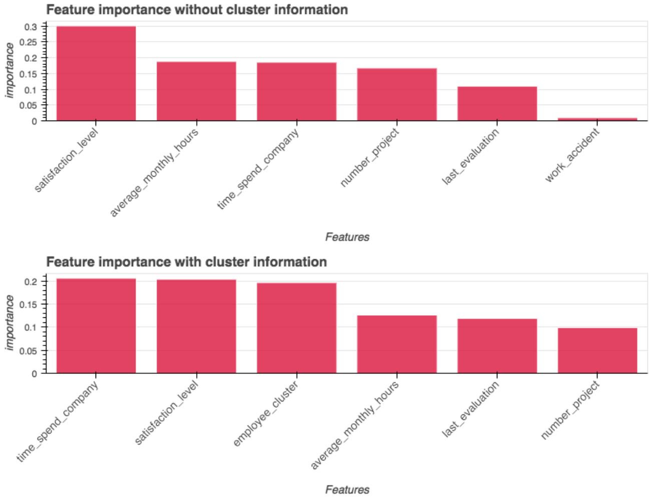
The additional information created through clusters are used in the prediction process and scores are evaluated compared to the scores without the cluster information. In order to get a better understanding on its performance all prediction algorithms are fed the new information.

The largest AUC score gain takes place for SVC followed by LinearSVC and LogisticRegression. The score gained for DecisionTreeClassifier is marginal as well as the loss for RandomForestClassifier. Based on cross-validation results the supervised classification has only produced slight changes in accuracy. RandomForestClassifier is still the leading algorithm for this problem (Table IX).

TABLE IX. AUC score after cluster-enhancement

Classifier	AUC	Time	+/- Score
RandomForestClassifier	**0.9897**	0.7376	-0.0007
DecisionTreeClassifier	0.9752	0.3140	-0.0013
SVC	0.9668	9.4367	0.0577
LogisticRegression	0.8209	0.2859	0.0030
LinearSVC	0.8203	1.6123	0.0030

FIG. XXII. Feature Importance for Leading Algorithm in Score with and without Cluster Information



1) Shift of feature importance: It might be interesting to see which features gained importance for the leading algorithm and compare it to the base results in the beginning. Fig. XXII indicates that the most important feature on the base prediction is satisfaction_level. After clustering the data and adding an additional discrete variable for all data points, feature importance and ranking change. As shown in the lower plot of Fig. XXII, the prediction process with cluster information satisfaction_level is still very relevant. Yet time_spend_company took over the first position, followed by satisfaction_level and employee_cluster. Overall through adding this additional variable, the ranking looks more balanced.

E. Parameter Tuning

In order to find the best parameter setup, a Grid Search is performed. The basic criteria for deciding on a set of parameters is its AUC scoring. In order to emulate a real-life scenario only 80% of the training data will be used. 20% of the data will serve as a sanity check to offer a better assumption on how this tuning would perform on new data. Similar as before a key focus is on stratifying the data to make sure the training and testing sets are having similar proportions of prediction cases. Parameters for all estimators and their values used to perform grid search can be seen in Tab. X.

After performing grid search and defining the best fitting parameters the training score of Random Forest improved to

TABLE X. Parameter and Values for Grid Search

Estimator	Values
LinearSVC	
C	range from [1, 11]
loss	hinge, squared, hinge
penalty	l2
max_iter	100, 200, 300
Logistic Regression	
C	range from [1, 11]
penalty	l2
solver	sag, newton-cg, lbfgs, liblinear
warm_start	True, False
Decision Tree	
criterion	gini, entropy
max_features	auto, sqrt, log2, None
max_depth	None, 2, 5, 10
SVC	
C	1, 2
kernel	poly, rbf, sigmoid
degree	2, 3, 4
Random Forest	
n_estimators	10
criterion	gini, entropy
bootstrap	False, True
max_features	auto, sqrt, log2, None
max_depth	None, 2, 5
min_samples_split	2, 4
warm_start	True, False
max_leaf_nodes	None, 4, 6

.9996 and the testing score scored .9773 which is remarkable since only 80% of the initial training data was used. To further validate this setup a 5-fold cross-validation is performed. Tab. XI shows all estimators and their final parameter setup. The average AUC score using the tuned Random Forest model is .9876 which is almost as high as the initial cross-validation score. Even though this is a slightly lower score, since it's a substantially smaller data set these parameters will be used further down the road.

VI. Results

In this section the results of the research are being discussed as well as additional thoughts on robustness and stability of the model.

A. Evaluation

The final leading algorithm is a RandomForestClassifier which has a good performance right from the base. The data set was exceptionally clean and there were only minor things to be adjusted. Overall research has shown that the complexity of employee turnover prediction is a multi-variate and highly complex task. Research suggests severe problems when it comes to generalization in this domain. Therefore this research has focused on cross-validation wherever possible. The final model seems reasonable and is aligning with solution expectations mentioned earlier. The final parameters of the model seem to leave enough room for generalization and a high performance on new data. Although the final model has been tested with various inputs to evaluate whether the model generalizes well to unseen data there is still a testing set available that will be used in this section to draw further conclusions about the model. But before heading to this final step there is the question of model stability and robustness that needs to be addressed.

TABLE XI. Result of Parameter Tuning and Prediction Score (AUC)

Estimator	Value
LinearSVC	
C	3
loss	squared_hinge
penalty	l2
max_iter	300
Prediction Score	**.8828**
Logistic Regression	
C	1
penalty	l2
solver	sag
warm_start	True
Prediction Score	**.8852**
Decision Tree	
criterion	gini
max_features	None*
max_depth	10
Prediction Score	**.973**
SVC	
C	2
kernel	rbf
degree	[non-applicable for rbf]
Prediction Score	**.9664**
Random Forest	
n_estimators	10
criterion	entropy
bootstrap	True
max_features	auto
max_depth	None**
min_samples_split	4
warm_start	True
max_leaf_nodes	None***
Prediction Score	**.9876**

* None means max_features equals the existing amount of features.

** None means nodes are expanded until all leaves are pure.

*** None means unlimited possible number of leaf nodes

B. Model Stability

To test the model for its stability beyond cross-validation a subset of the testing data is being created that contains only a small shuffled sample including 40% of the training data and additional 20% for testing purposes. If robust, the model is supposed to have a similar predictive outcome for training and testing on the stability set. On a smaller subsample of 40% the prediction algorithm still works very well. Its score on the testing score even increases up to .993 indicating a stable model which was to be expected by the usage of cross-validation down the road.

C. Testing

As a final step, all models will be used on the test set. This will be the final determinator on which algorithm and model setup performs the best. The following setups will be tested for LinearSVC, LogisticRegression, SVC, DecisionTreeClassifier and RandomForestClassifier:

- Base model, with little to no tuning to the parameters
- Clustering-enhanced data, with little to no tuning to the parameters

- Clustering-enhanced data, with tuned parameters

The scoring method will be AUC in order to be able to compare the scoring to mentioned research by [8].

TABLE XII. AUC SCORES FOR BASE, CLUSTER AND TUNED PREDICTION

	Base	Cluster	Tuned
RandomForestClassifier	0.9776	0.9770	**0.9805**
DecisionTreeClassifier	0.9702	0.9711	0.9568
SVC	0.7391	0.8313	0.8313
LogisticRegression	0.6409	0.6466	0.6501
LinearSVC	0.5992	0.6155	0.6367

Using AUC as a metric the highest scoring model is a tuned RandomForestClassifier trained on a cluster-enhanced data set and achieves a score of .9805 on the hold-out testing set.

D. Precision vs Recall

As mentioned earlier, one problem of AUC can be low recall scores. Some classifiers can have a low recall but a very high AUC score. In order to protect the established model against low recall, F1 scoring is implemented. For F1 scores to be high, both precision and recall have to be high.

TABLE XIII. F1 VS AUC SCORING ON TESTING SET

	F1	AUC	Delta
RandomForestClassifier	0.9773	0.9805	-0.0032
LinearSVC	0.4378	0.6367	-0.1989
LogisticRegression	0.4644	0.6501	-0.1857
SVC	0.7895	0.8313	-0.0418
DecisionTreeClassifier	0.9483	0.9711	-0.0228

Tab. XIII shows the prediction result when using 5-fold cross-validation and f1 scoring with all tuned estimators. When looking at the F1 test scores for the leading algorithm, there is only a small delta between the AUC score and the F1 test results. For LogisticRegression and LinearSVC the delta is quite large though. Using F1 scores put these scores below .5 which makes them worse than random guessing. It can be said that the established model scores high on both, precision and recall.

E. Justification

This research achieves even higher results without the need for more computational expensive algorithms such as Extreme Gradient Boosting. [8] achieved a .86 AUC score on a hold-out data as a best result. In this research we presented a method to achieve good prediction results with the combination of unsupervised and supervised learning algorithms. The resulting method achieves a .9805 AUC score and is over .12 higher than the research result presented in [8].

The reason for this might be hard to determine without access to the original data used in [8]. Based on the data set of this research one possibility could be the clean variables that include only little noise. While [8] had to combine information from multiple sources and work around a quite noisy data set this research had the benefit of working with exceptionally little data infused obstacles. Therefore the higher results must also be linked to the data set. As almost always in machine learning, better data beats better algorithms. Based on the problem statement and the task that was at hand this research can be seen as a successful solution though.

VII. Conclusion

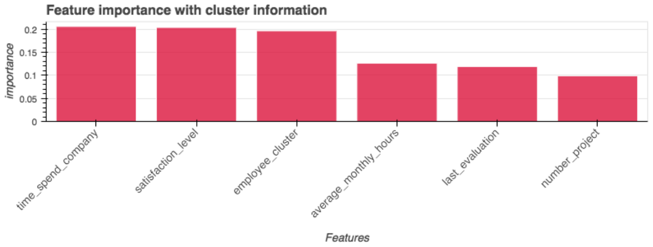
This section will first focus on important qualities of this research and interesting findings, before starting an in-depth reflection about the process and its key elements.

A. Feature Importance and pro-active behaviour

As discussed earlier, for most companies it is as important to understand why an employee is about to leave or leaving as it is to predict. In order to act in advance and possibly change the circumstances in a more favorable way for top performers and valuable assets the variables that are most important to employee turnover should be examined. This research offers a way of looking into the importance of certain aspects of work culture to predict employee turnover.

Based on the model presented by this research following categories should be examined strongly per individual to make sure to act before it is too late:

FIG. XXIII. Feature Importance with Cluster Information



- Front-runner is time_spend_company suggesting employee turnover to be heavily influenced by the time an employee was with a company. The importance supports the finding that work relationships nowadays tend to have a shorter life cycle. Looking at the initial remarks about sinking tenure in the work force, this seems to be just fair. A way of anticipating this fact is opening up pathways that lead to a diverse work environment filled with new tasks and possibilities for professional growth. If the company is big enough, encouraging employees to work on their personal skill set or even promoting different possibilities within the companies might be a good way to achieve a higher retention rate.
- Next up is satisfaction_level. There are many ways to measure this in a work space. For any company to capture this rather abstract value in a correct way it has to be in touch with its employees. There is only so

much information you can gain by looking at business numbers. Measuring this variable seems to be crucial for employee turnover. A finding that is supported by management and psychology studies mentioned earlier as well.

- Third is employee_cluster. As mentioned earlier, employee_cluster can be very helpful when it comes to categorizing top performers but also employees at risk. With additional continuous variables this could help put employees on the map in even very complex work domains.
- average_monthly_hours can be an indicator for both, overworked employees or employees that are not being challenged sufficiently.
- Rating employees and using their last_evaluation might be a good indicator but even better would be time series information on how their performance has changed over time. In addition there are multiple layers when it comes to evaluations on the job. Having a multifaceted evaluation score sheet that's focusing on performance but also on social skills could be a great indicator for employee turnover.
- Lastly work responsibility can be measured by number of projects. It seems as if empowering people helps keep them on board.

Given that work performance is an ever-changing subject, time series analysis on all mentioned features could offer additional insights.

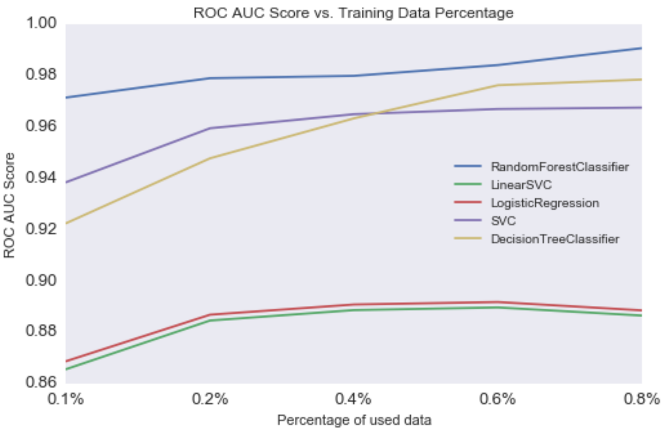
B. *Scoring with increasing training data*

One important characteristic for robust models is the behavior of their prediction accuracy when using it on different amounts of training data. For this exercise a range between 10% and 80% of initial training data will be used and the average scoring on a 5-fold cross-validation is being measured for all algorithms used in this research. The results can be seen in Fig XXIV. An interesting observation can be made when looking at cross-validation scores based on the amount of training data available. Even with as little as 10%, tree-based algorithms work very well on the data. Which speaks to their robustness for this specific task. The biggest score gain can be observed for DT, which pops from around .92 to a score just shy of .98. Both Logistic Regression as well as LinearSVC are on lower scoring levels, improve only slightly and perform almost as bad on 10% data as soon as they reach 80%. This could be linked to overfitting due to parameter tuning.

C. *Reflection*

1) Introducing problem, domain and a potential solution: The process developed in this research took a raw data set filled with employee information and turned it into a predictive analysis regarding employee turnover. The research was split into several parts. It starts with an introduction that states additional background information and gives an overview about recent research about employee turnover. In this part the problem and a possible solution is also stated and based on the literature review a benchmark has been established to measure the quality of this research paper. An interesting takeaway from this part was the complexity of the domain and various multi-layered approaches that have been executed in the search of workable solutions.

FIG. XXIV. AUC SCORE WITH INCREASING AMOUNT OF TRAINING DATA



2) Exploring the data and its peculiarities: Afterwards an in-depth analysis of the data took place. Starting with data exploration and steps to deal with individual traits of the data set. From univariate analysis of continuous and discrete variables to multivariate perspectives including but not limited to regression and correlation analysis. This section highlighted the importance of context, when it comes to a human related topic. Just as important as creating a prediction algorithm is understanding the context and inner relationships of its variables. The takeaway of this segment were multiple possible employee stories and input for additional measures when it comes to employee turnover.

3) Thoughts on models and algorithms: After establishing theoretical and topical groundwork several avenues for machine learning solutions were explored. Starting with a collection of supervised and unsupervised algorithms that could potentially lead to a successful solution and their intuition were mentioned. This segment tried to give a brief introduction into the criterias and upsides which led to a set of algorithms. Afterwards the benchmark was theoretical introduced and explained in deeper detail. A major takeaway was the complexity of combining supervised and unsupervised sets especially when dealing with a mix of variable classes. Since it's not that easy to apply clustering algorithms to a mix of continuous and discrete variables a compromise had to be found. Another important takeaway was the discussion about prediction scoring. Adding F1 scoring to the mix increased the reliability of the results.

4) Exploring methods for handling the data: In the next segment the methodology behind handling the data and dealing with its detected peculiarities was discussed. From data preprocessing with steps such as missing value or outlier detection and handling as well as One-Hot encoding for categorical variables to feature normalization and the calculation of feature relevance.

5) Implementing a solution: The implementation phase was split into three prediction and two refinement runs. Combining efforts of unsupervised and supervised learning algorithms this segment tried to put into action what was introduced in the solution statement. Before starting with the first prediction, 20% of the data was set aside for testing. For all supervised learning algorithms stratified 5-fold cross-validation was performed in an attempt to create an algorithm that would generalize well on new data. First it started with a basic implementation of the supervised learning algorithms, measuring their success based on the same score as the benchmark. Afterwards an unsupervised learning algorithm was implemented to find clusters and additional relationships in the data. The cluster labels were added to the data set and another run of prediction took place. This was a difficult task and prior to starting the research I hoped the information available was richer when it comes to clustering. Because of the small amount of continuous variables and the complexity of mixed clustering methods the clustering results are suboptimal and led to only minor changes in the prediction. In an attempt to fine tune the algorithms, grid search was performed on all supervised algorithms to find a set of parameters that worked the best based on its cross-validation AUC score. Yet, even with slightly reduced training scores on the clustering data set, the final model performing best on the hold-out set was a combination of tuning and cluster-enhancement.

6) Additional model testing: Afterwards robustness of the model was tested by performing additional cross-validation runs on smaller subsets of the training.

7) Success of the model: The final model and solution does fit the expectation stated in the beginning. Yet it is important to point out that this solution is tied to the present data set. With a data set that is more noisy the problem of generalization as mentioned in some of the presented literature may occur especially when dealing with tree-based algorithms that don't implement pruning. Cross-validation with pruning parameters and boosting are some of the options available to work against this problem. The solution presented here could be used in a general setting to determine not only employee turnover but also feature relevance in companies with a certain size and reporting structure.

D. Improvement

There are a few suggestions to improve this research that would go beyond the scope of this analysis.

1) More and more diverse data: The available information for this research has its limitations. Although the data was nice to handle and exceptionally clean from the beginning there were some aspects that felt could use improvements. In an attempt to get more insights a follow up research could start pulling additional information from a variety of sources, gather more granular splits on some of the existing variables or try to reverse engineer some information that might be hidden in the data set. However with missing code-book and contextual information for the data it was quite hard to add additional sources since most of the available information is highly regional and to a certain extend company specific.

2) Additional algorithms: For similar research more sophisticated algorithms could be used. There have been great results in this domain with Extreme Gradiant Boosting models. In addition the implemenation of an artificial neural net might be possible if the data set grows to an extend where it is feasible to train a net.

3) Computational Efficiency: Some of the parts in this research are computationally expensive. Especially computing the silhouette score and performing grid search on all of the algorithms with a broad variety of parameter settings can take a long time to process. An implementation of this research should focus on parallel computing which is not only possible for specific segments but also for tree-based algorithms such as Random Forest. Another option is preprocessing some of the data upfront and establishing a knowledge data base that can be loaded rather than computed every time in order to minimize computing time.

REFERENCES

[1] Allen, D. G., Moffit, K. R., & Weeks. K.P, *Turnover intentions and voluntary turnover: The moderating roles of self-monitoring, locus of control, proactive personality, and risk aversion.* Journal of Applied Psychology, Vol 90, 980990. 2005.

[2] Cotton, J. L. & Tuttle, J. M., *Employee Turnover: A Meta-Analysis and Review with Implications for Research* Academy of Management Review, Vol 11, 5570. 1986.

[3] Hand J. D. & Krzanowski J. W. *ROC Curves for Continuous Data* CRC Press, p. 26, 2009.

[4] Holtom, B., Mitchell, T., Lee, T., & Eberly M., *Turnover and retention research: A glance at the past, a closer review of the present, and a venture into the future* Academy of Management Annals, 2: 231-274, 2008.

[5] Michaels, C. E. & Spector, P. E., *Causes of employee turnover: A test of the Mobley, Griffeth, Hand, and Meglino model.* Journal of Applied Psychology, Vol 67(1), 53-59. 1982.

[6] Mitchel, J. O, *The effect of intentions, tenure, personal, and organizational variables on managerial turnover.* Academy of Management Journal, Vol 24, 742-750. 1981.

[7] Mobley, W. H., *Intermediate linkages in the relationship between job satisfaction and employee turnover.* Journal of Applied Psychology, Vol 62(2), 237-240. 1977.

[8] Punnose, R. & Pankaj A., *Prediction of Employee Turnover in Organizations using Machine Learning Algorithms* International Journal of Advanced Research in Artificial Intelligence, Vol 5. 2016.

[9] Kim, H., *Quora*, https://www.quora.com/What-is-an-example-of-real-world-application-of-Gaussian-Mixture-Models/answer/Hongsun-Kim. 2015.

[10] Trevor, C. O, *Interactions among actual ease-of-movement determinants and job satisfaction in the prediction of voluntary turnover.* Academy of Management Journal, Vol 44, 621638. 2001.

[11] United States Department of Labor - Bureau of Labor Statistics, *Employee Tenure Summary* https://www.bls.gov/news.release/tenure.nr0.htm, September 22, 2016.

[12] Weisbeck, D., *Fact or Hype: Do Predictive Workforce Analytics Actually Work?* visier.com. 2015.