Employee Attrition Analysis

IBM HR Data

Contents

Project: IBM HR – Employee Attrition

Summary Overview:

The goal of this report is to analyze employee attrition. Attrition is described as the gradual loss of employees over time. Attrition is a major issue for all organizations, where it can lead to implications in staffing, employee morale, project costs, loss of experience, and a general hindrance to organizational growth. We will examine the most important factors that influence attrition within an organization. We will consider if these factors are within the control of the organization and what actions can used to mitigate or combat attrition. We will also analyze current trends in HR and how these apply to our analysis, where finally, based on our results, we will conclude with insights and recommendations.

Introduction:

Employee capital is one of the greatest assets an organization can possess. Companies can spend *as much as 70% of total business costs* on employees. These costs include salaries, training, recruitment, and skill investments. Furthermore, recruiting and keeping top talent is important to the growth and long-term viability of any company. Often employees hold key characteristics that are instrumental in moving the company forward. Knowing this, when employees decide to quit or leave a company, it can be a serious issue. With each employee, the company loses its direct investment along with all the knowledge and experience that the employee would have inherently provided. In the field of Human Resources, HR, when employees decide to quit, this is referred to as employee attrition, and this is the focus of our analysis.

In looking further into the causes of attrition. We see that there is overlapping evidence for why employees decide to leave, where in most cases it is due to the following reasons: unsatisfying compensation, unsatisfactory benefits, lack of growth or development opportunities, issues with work-life balance, poor management, poor work conditions, and lack of recognition for work accomplishments or value add in the workplace.

In this report we will be discussing the validity of these reasons through a general analysis of the publicly available IBM employee dataset. We will analyze the dataset through a generally understood set of procedures. These procedures are as follows: research, exploratory data analysis, data manipulation and cleaning, model preparation, model building and evaluation, and data insights and analysis. This process can be approached procedurally or iteratively. As we will discuss, we iterated between data manipulation, model preparation, model building, and the analysis thereafter.

Overview of Data

The core of our analysis is derived through a publicly available online dataset, providing variables akin to what typical Human Resources personal would have at available. This dataset is known on Kaggle as the *IBM HR Analytics Employee Attrition & Performance* dataset. It is comprised of over one thousand and four hundred observations and thirty-five features—features and variables is used interchangeably in this report to represent the columns for which observations pertain to.

Within the dataset we have a mix of numeric and categorical datatypes. The initial dataset contains twenty-six numeric variables and it contains nine categorical data types. Examples of numeric data types include employee age, their monthly income, and the years that they've been working with the company. Examples of available categorical variables include education, gender, job role, or the department that the employee works with.

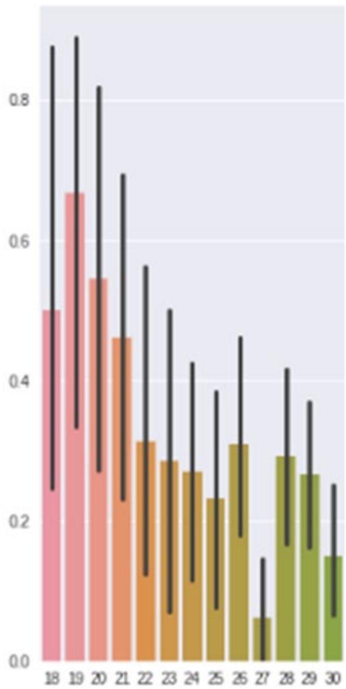
It is important to verify the construction of the dataset and its variable types. This is because the make-up of the data directly influences the nature of what is possible in the analysis process. If we put aside data manipulation and management techniques—which we will discuss later in the report— each model is best suited for certain data types. If we take linear regression as an example, it is best suited for numerical input, where it then predicts a linear relationship on a numeric dependent variable. On the other hand, if we utilize a model like Naive Bayes or a process like analysis of variance, then these two options are better suited to categorical independent variables. As we will find out later in the report, we will use inherent programming options to manipulate our data to work with the models that we have selected.

On the topic of model selection, the goal of our analysis is to define attributes associated with employee attrition. It is also to generate insights for which a Human Resources department can then take actions that would mitigate attrition. The goal of our analysis is not to generate a highly accurate prediction model; however, we will include several prediction modelling techniques, like Random forest, Support Vector Machine, and Gradient Boosting. We hypothesize that decision tree-based modeling will work best for our analysis, as it works well with the available data; furthermore, the pruned and tailored out that is generated by a decision tree model is easier to understand than that of other models.

List of features:

Before discussing exploratory data analysis and modeling, we will briefly present a list of the variables within the dataset and either their pertinent attributes or the value range they contain. The full list is as follows: Age, 18-60; Attrition, Yes or No; BusinessTravel, frequently, rarely, none; DailyRate; Department, sales, HR, R&D; DistanceFromHome, 1 to 29 miles; Education, below college, college, undergrad, masters, doctorate; EducationField, HR, technical, marketing, life sciences, medical, other; EmployeeCount; EmployeeNumber; EnvironmentSatisfaction, low, med, high, very high; Gender; HourlyRate; JobInvolvement, low, medium, high, very high; JobLevel, 1-5; JobRole, sales rep, lab technician, HR, sales exec, research scientist, manufacturing director, heathcare rep, manager, research director; JobSatisfaction, low, med, high, very high; MaritalStatus, married, divorced; MonthlyIncome; MonthlyRate; NumCompaniesWorked, 0-9; Over18; OverTime; PercentSalaryHike 11-25; PerformanceRating 3-4; RelationshipSatisfaction low, med, high, very high; StandardHours 80; StockOptionLevel 0-3; TotalWorkingYears 0-40; TrainingTimesLastYear 0-6; WorkLifeBalance bad, good, better, best; YearsAtCompany 0-40; YearsInCurrentRole 0-18; YearsSinceLastPromotion 0-15; YearsWithCurrManager 0-17.

Exploratory Data Analysis and Discovery



Age v Attrition

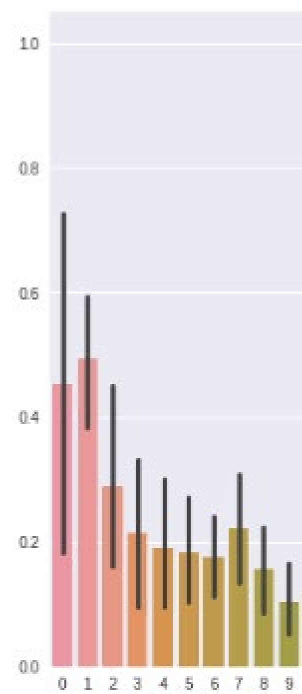
When analyzing a dataset, it is important to initially explore the constructs of the data, getting a sense for how the content within each observation varies, the ranges, the most influential variables, , the least. The simplest way to approach exploratory data analysis is through visualizations and basic statistical analysis. In the section below, we will discuss the exploratory data analysis process on each variable, the resulting connection to attrition, and our general consensus towards the importance and use of the respective variable.

To the right we see a cropped section of a graph that illustrates the attrition rate along with an employee's age. We see how between the ages of 18 and 21, there is a major increase in attrition, while as the years continue, and especially looking at age 27, the attrition rate is significantly lower. This same pattern is apparent when looking at the years worked by employees. We notice that the first three years of employment incur the highest attrition rates.

Below we will briefly discuss the variables available to us through the dataset. We will also consider how each variable relates to the attrition rate and any meaningful bi-variate or multi variate analysis. We can begin by examining demographic data, including age, gender, marital status, academic field of study, and education level. The age range within this dataset is between 18 to 60. This gives a full age range from completely green—employees that have no work experience—to employees who are ready to retire. The marital status includes whether an employee is single, married, or reportedly divorced.

Education includes five levels, from no college experience to having a masters or doctorate degree. We see that the variance in attrition is not large when comparing employees based on their degree. Having no college experience is corresponds to a 52% attrition rate, while having a masters or doctorate degree roughly corresponds to a 45% attrition rate. We do see a slightly higher increase in age per the degree, where an employee with no college experience has an average age of 32, while employees that have a Masters or Doctorate both have an average age of 39 years. All education levels also have similar average years at the company, where employees with no college experience have an average of 6.5 years at the company, and those with a higher degree have between 7.5 and 8.3 years. This information on education leads us to believe that education level does not have a significant impact on the attrition rate. If there was a greater variance in the years worked or the age group, then we would expect a greater impact, especially if they average down at the lower age group where attrition rates are highest.

Years at Comp V Attr.



When analyzing marital status, we do see a greater proportion of single employees leaving the company versus those that are married or divorced. We see that on average, married and divorced employees have an attrition rate of 40%, where employees that indicated being single, have an attrition rate of 64%.

Concerning environment satisfaction, we do see a positive correlation between a reportedly low environmental satisfaction versus a medium, high, and very high. We see that the average attrition rate for medium, high, and very high is 14%. The attrition rate jumps to 25% with a environmental rating of low.

Looking at job involvement, we see that there is also a high correlation to attrition rate. Employees who maintain a job involvement rating of 1, or low, have an attrition ratio of 33.5%. This is a stark contrast to those who have a very high job involvement rating, where the attrition rate hovers around 9%. If we average job involvement, combining medium, high, and very high, we see an attrition ratio of 13.6%. This is roughly two and a half times lower than employees with a low job involvement rating. Additionally, after oversampling for this variable, we found that a low job involvement corresponded to an attrition rate of 70%. This is a significant finding.

Job satisfaction is generally understood to be an important factor on employee attrition. After examining for this variable, we see that the range between employees who report a low job satisfaction and those that report a very high job satisfaction, is 22.8% versus 11.3 %. This is approximately double and is notably concerning, although it is not as drastically different as we might have expected. It is also important to note that while this variable is inherently indicative of an employee's sentiment towards the job, it is also very difficult to control for. There can be an array of circumstances that affects an employee's degree of satisfaction where, furthermore, while certain circumstances may effect one employee drastically, it may effect another employee very little.

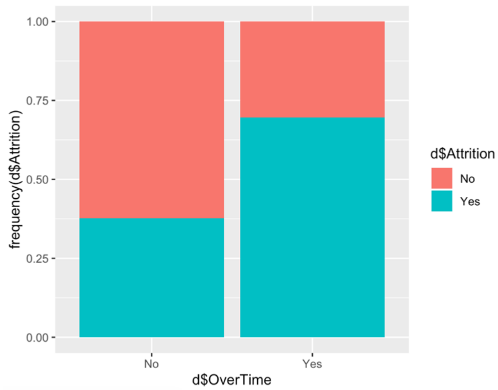
Performance rating in context to this dataset is not notably correlated to attrition. This seems mainly due to the fact that we do not have performance ratings of low and good available. The dataset only contains employees with a performance rating of excellent and outstanding. Between these two instances, we see the same attrition rate of 16%. We would note that while performance rating is typically a very good indicator of how an employee fits within a company, for our analysis, it is not practical to use within our analysis.

We see that relationship satisfaction has a notable, though ambiguous connection to attrition rate. The attrition rate between when an employee who indicates a satisfaction of low versus very high, is different by 9%. In comparison to job involvement, where there is a difference of 24% between the low and high end. Relationship status is concerning, yet to a lesser degree. By looking deeper at these two variables, we do find interdependent differences. For instance, when relationship satisfaction is low and job involvement is low, the attrition rate is 23.5%. When relationship satisfaction is low and job involvement is high, the attrition rate drops to 7%. When relationship satisfaction is very high and job involvement is low, the attrition rate is 43.5%. But when relationship satisfaction is high and job involvement is also high, the attrition rate is 12.5%. While we do notice this relationship, after reviewing the statistical correlation between the two variables, we see that it is only 0.03, which is very low. Additionally, relationship status is another variable that is difficult to control for, since there are many factors that go into determining an employee's degree of satisfaction towards their manager or colleagues.

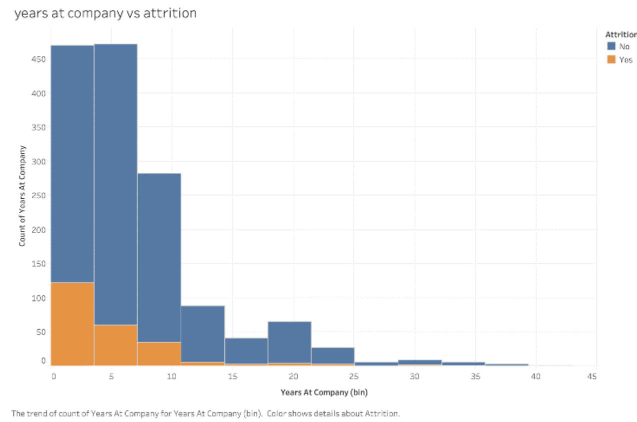
Lastly, we examined work-life balance, since this is presumably an important factor in modern day company culture. This variable carries a rating of between 1 and 4, where 1 is equivalent to bad work-life balance and a 4 is equivalent to the best work-life balance. Looking at the data, we find that the attrition rate when work-life balance is bad, is 31.25%; furthermore, on average, when work-life balance is good, better, or best, the attrition rate is 16.2%. This is approximately a two-fold difference.

When reviewing the statistical correlation between work-life balance and attrition, we find that there is again, a very low correlation of 0.064. For this reason, and due to the fact that the dataset does not state what factors lead to the a bad or a good work-life balance, we chose not to utilize this variable in later portions of our analysis.

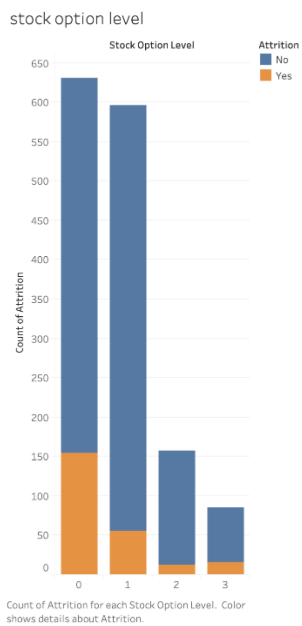
In the figures below, we included attrition rates as they correspond to several key variables. These include years at the company, overtime, stock option level, and lastly, stock option level in conjunction with overtime. We see in the first figure that when overtime is clocked in by employees, that the attrition rate is nearly 75%, while the attrition rate is 37% when they do not. We can see in the second figure that attrition rates are exceedingly high from when an employee start with the company at year zero, to the tenth year. This proportion is much higher within the first three years. If we look at stock option level in conjunction to overtime, we see that when stock option level is not offered –equal to zero—and overtime is clocked in, that the attrition rates are above 80%. The pattern is successively high even while stock options increase. When overtime is not clocked in and when stock options are not given, the attrition rate is high at 54%, which is expected due to correlation to no stock options, but as stock options are increase and while overtime is equal to no, then the proportion of attrition rates begin to drop. We see that stock option level two is the best alternative, with an attrition rate of 26%, nearly half that of an employee without stock options.



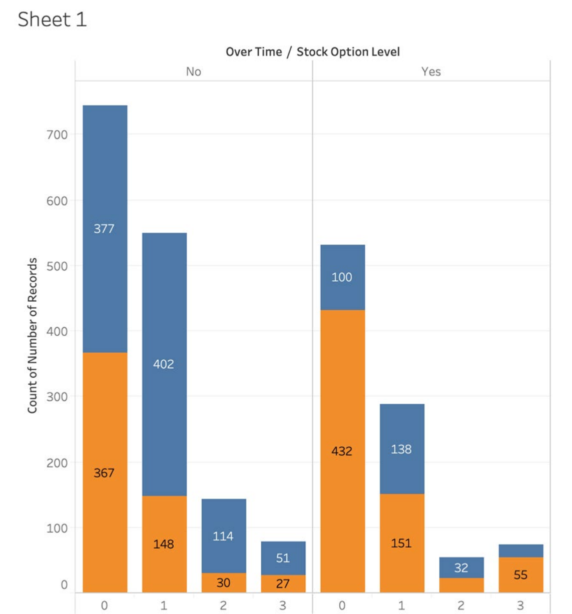
Overtime vs Attrition



Years at company vs Attrition



Stock option level vs Attrition

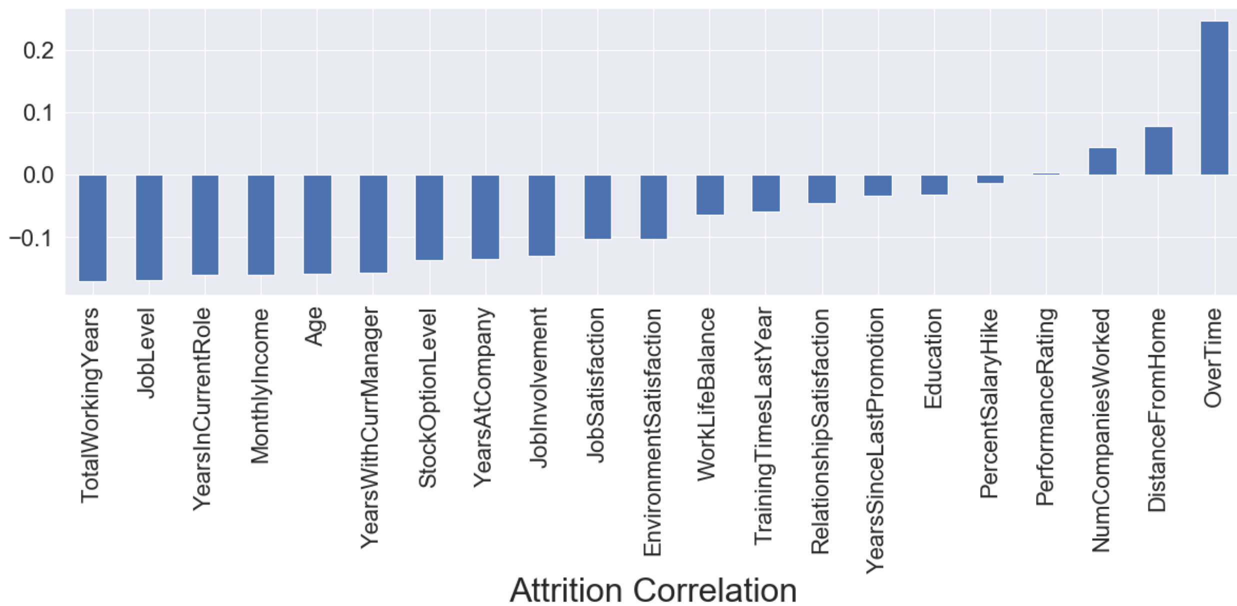


Overtime and Stock Option vs Attrition

Dependent Correlation

After reviewing the correlation between variables and attrition, we see that the average correlation for attrition is -0.0696. The lowest correlated variable is an employee's total working years, with a correlation of -0.171. The variable with the highest correlation is overtime, with a correlation of 0.246. Correlation values that we positive indicate that they are positively correlated with attrition rates—more overtime is correlation with higher attrition rates—while correlation values that we negative, correlate with lower attrition rates. This means that overtime, along with distance from home, and number of companies worked, are positively correlated to an employ leaving. Variables like total working years, job level, monthly income, and age are negatively correlated to attrition rates.

	Attrition
OverTime	0.246118
DistanceFromHome	0.077924
NumCompaniesWorked	0.043494
PerformanceRating	0.002889
PercentSalaryHike	-0.013478
Education	-0.031373
YearsSinceLastPromotion	-0.033019
RelationshipSatisfaction	-0.045872
TrainingTimesLastYear	-0.059478
WorkLifeBalance	-0.063939
EnvironmentSatisfaction	-0.103369
JobSatisfaction	-0.103481
JobInvolvement	-0.130016
YearsAtCompany	-0.134392
StockOptionLevel	-0.137145
YearsWithCurrManager	-0.156199
Age	-0.159205
MonthlyIncome	-0.159840
YearsInCurrentRole	-0.160545
JobLevel	-0.169105
TotalWorkingYears	-0.171063



Data preprocessing

Data preprocessing is necessary prior to running modeling due to constrains in the modeling process. For instance, certain models cannot process categorical data. For this we need to convert the categorical information within a variable into a numerical representation. The method we instill for this process is attribute expansion, where we use a binary value of one or zero to represent the appearance or absence of a value.

```
def transform(row):
    if row['PercentSalaryHike'] < 15:
        val = 'Yes'
    else:
        val = 'No'
    return val

employees['low_percentage_hike'] = employees.apply(transform, axis=1)
employees = employees.drop('PercentSalaryHike', axis=1)
```

```
def transform_relation_sat(row):
    if row['RelationshipSatisfaction'] == 1:
        val = 'Yes'
    else:
        val = 'No'
    return val

employees['low_relationship_satisfaction'] = employees.apply(transform_relation_sat, axis=1)
employees = employees.drop('RelationshipSatisfaction', axis=1)
```

We transformed following variables using this method:

- Percent salary hike => Low_salary_hike (True, False)
- Relationship satisfaction => Low_relationship_satisfaction (True, False)
- Business Travel => Frequent Travel (Yes, no)
- Job Involvement => Low job involvement (Yes, no)
- Work-life balance => Low work-life balance (Yes, no)

We used the above methods because the yes categories had a very large percentage of attrition compared to other categories.

Our dataset had imbalanced data with 16% instances for class 'Yes' and 84% for 'No'. Because precision of predicting the higher number of 'Yes' classes is more important for this case, we used oversampling to balance both the classes.

In addition to categorical to numeric conversion, if is also necessary to declare a set of variables by degree of importance. For the dataset that we are using, which contains over thirty-five variables, we must certainly run feature select to define the most pertinent features to utilize.

After balancing the data, we used feature selection to identify the top features since we wanted to reduce the number of variables in our dataset. We used SelectKBest function of sklearn feature_selection library to identify top 20 features out of 35 variables in our dataset. Below we can see the output of this process. Note that monthly income, years at company, age, distance from home, and overtime are among the top selected features.

	Specs	Score
2	MonthlyIncome	411536.225257
5	YearsAtCompany	433.389238
0	Age	306.601455
1	DistanceFromHome	168.847410
13	OverTime	145.667368
14	StockOptionLevel	90.301831
12	MaritalStatus	52.232840
19	low_worklife_balance	31.705882
18	low_job_involvement	31.053763
17	frequent_travel	30.952381
9	EnvironmentSatisfaction	22.861395
3	NumCompaniesWorked	20.701826
11	JobSatisfaction	17.644864
16	low_relationship_satisfaction	10.695312
4	TrainingTimesLastYear	6.383004
6	Department	4.292568
10	Gender	1.063830
8	EducationField	0.805780
7	Education	0.453708
15	low_percentage_hike	0.326425

Gain Ratio feature evaluator

Ranked attributes:

0.0736	15	OverTime
0.0562	13	MonthlyIncome
0.046	1	Age
0.0445	18	StockOptionLevel
0.0301	21	YearsAtCompany
0.0266	20	WorkLifeBalance
0.0261	12	MaritalStatus
0.0224	8	EnvironmentSatisfaction
0.0189	3	BusinessTravel
0.0144	14	NumCompaniesWorked

Selected attributes: 15,13,1,18,21,20,12,8,3,14 : 10

Information Gain Ranking Filter

Ranked attributes:

0.1139	13	MonthlyIncome
0.0707	15	OverTime
0.0671	1	Age
0.0669	21	YearsAtCompany
0.0563	18	StockOptionLevel
0.0395	12	MaritalStatus
0.0273	14	NumCompaniesWorked
0.0213	3	BusinessTravel
0.0205	10	JobInvolvement
0.0185	5	DistanceFromHome

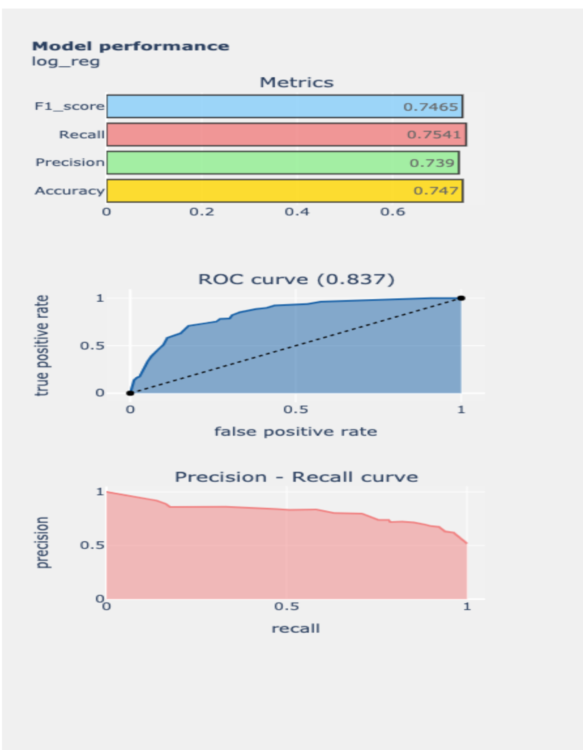
Selected attributes: 13,15,1,21,18,12,14,3,10,5 : 10

Note that monthly income, years at company, age, distance from home, and overtime are among the top selected features. There is crossover using information gain where the top features are monthly income, overtime, age, years at the company, and stock option level. Lastly, there is also cross over with gain ratio, which rates overtime as the highest, followed by monthly income, age, stock option level, and years at the company. For interpretability, we mainly focused on the gain ratio feature selection rankings.

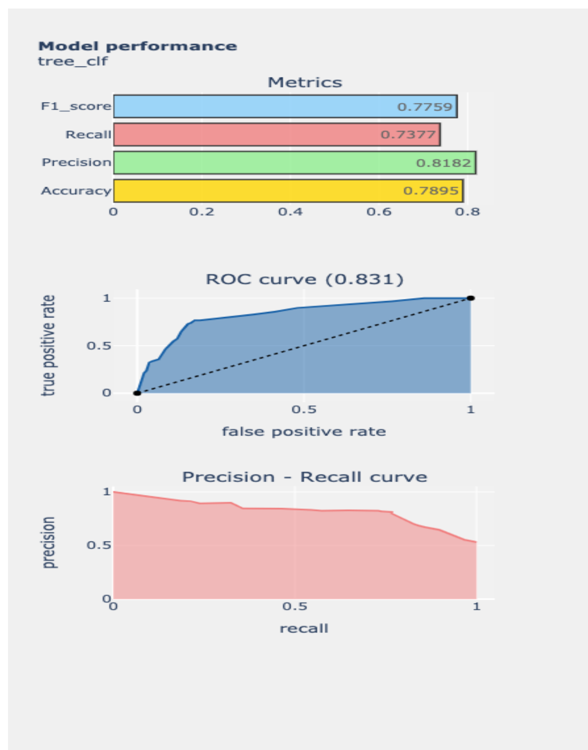
Modeling

After we did the data preprocessing, we proceeded with the prediction process. We decided to go with logistic regression and decision tree for modeling as the results for these two are interpretable. But accuracy and precision were pretty low for both the models. So, we tried random forest and gradient boosting too to improve the accuracy and precision. For initial modeling, we used all variables for prediction and found the top features for each classifier.

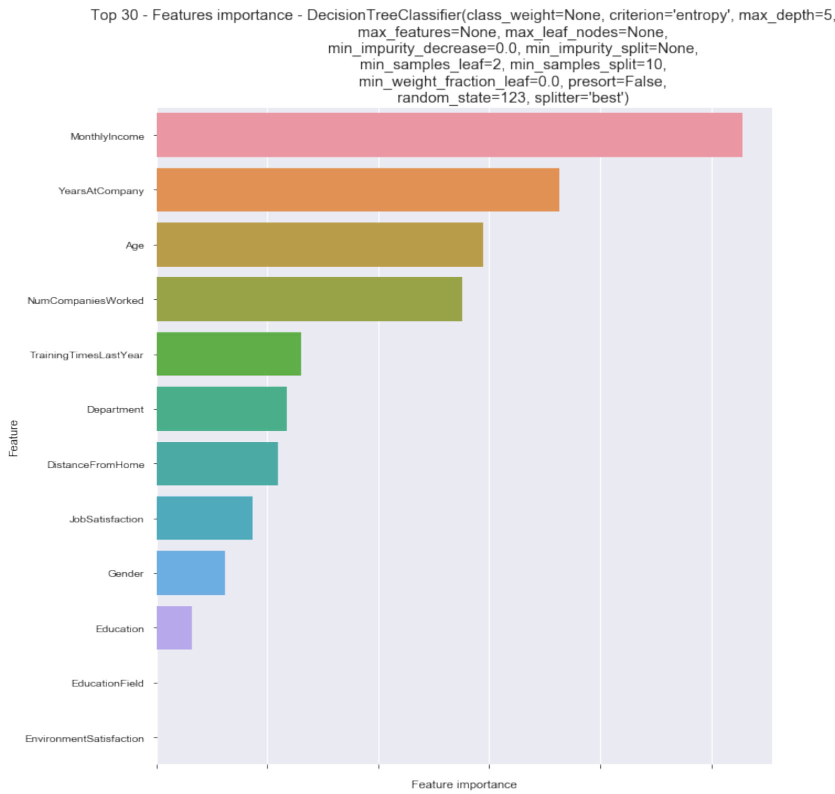
Logistic Regression results:

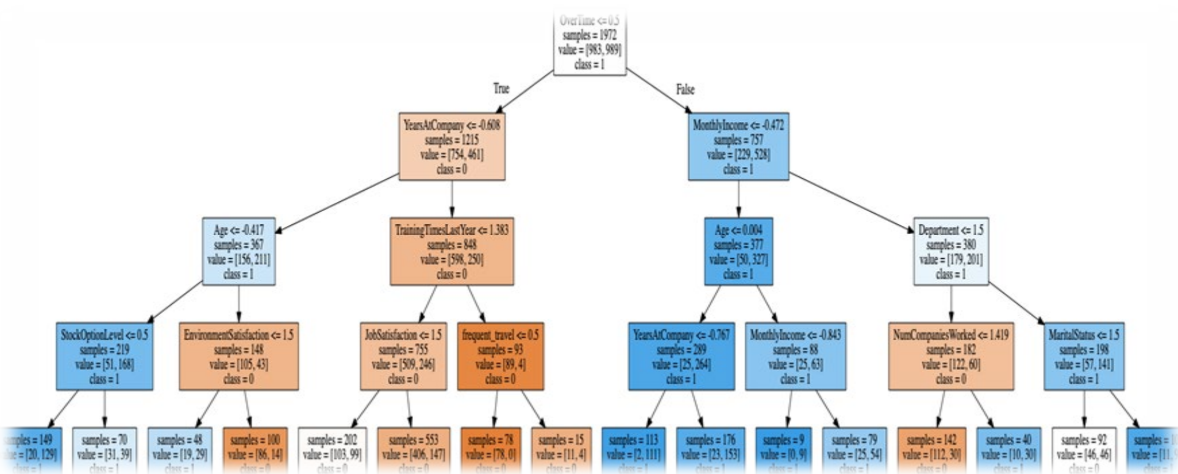


Decision tree results:



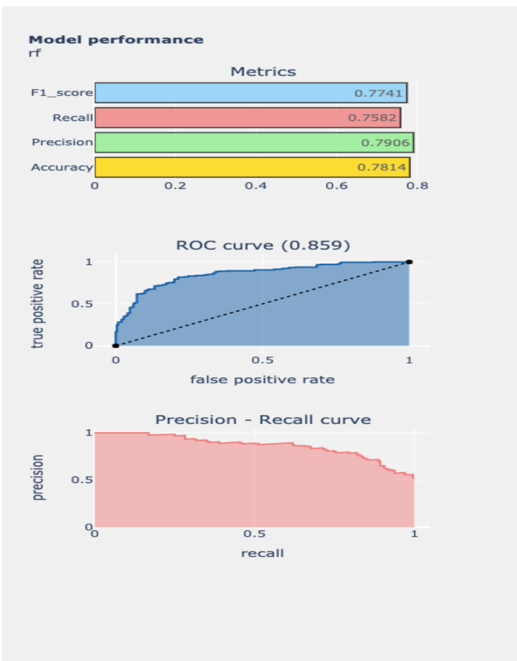
Decision Tree

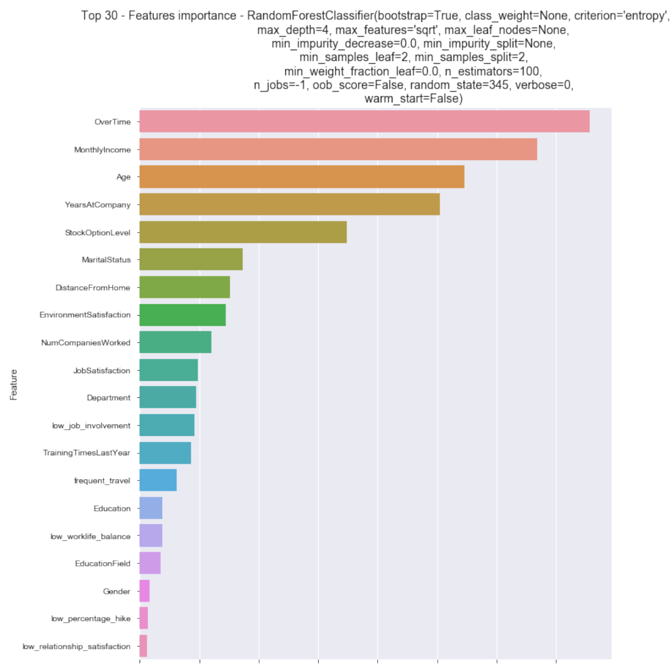




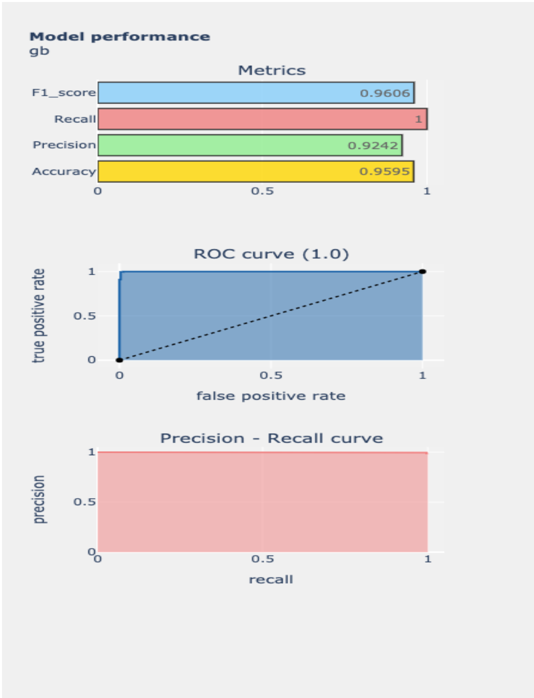
The above decision tree illustration can be utilized by HR personal in quickly determining key attribute string that lead to attrition. For example, if an employee has clocked in overtime, has worked for only one year, their training is less two times in a year, and they do not travel frequently, then class equal zero, or they are more likely to stay.

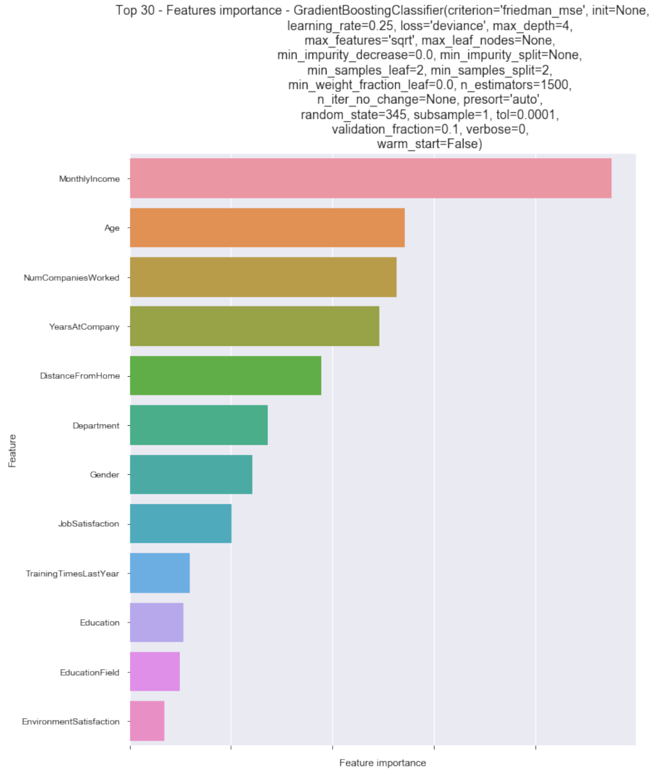
Random forest results:





Gradient boosting results





Based on the above modeling results, gradient boosting was certainly the top performer for our dataset. After identifying the top features from feature selection and feature importance of each classifier, we decided to choose top 12 features from our feature selection and ran the models again. The precision and accuracy were as follows:

Model	Accuracy	Precision
Logistic Regression	69.4%	67%
Decision Tree	75.5%	78.3%

Random Forest	75.9%	77.7%
Gradient Boosting	92.9%	87.4%

The accuracy and precision decreased slightly for almost all the models. But there wasn't much difference even though we reduced the number of attributes from 35 to 12 which meant that more than half of our variables were insignificant when it came to predicting the attrition. Monthly income, age, years at company some of the top variables in prediction for all the models.

Insights & Recommendations

Below we will discuss insights from the above analysis and recommendations that can result in action. To reiterate what we discovered through our above analysis, we found that among the most important attributes in determining attrition rates are overtime, monthly income, age, stock option level, and years at the company.

Let us briefly review monthly income. Research shows that it is related to satisfactory pay, where if an employee's monthly income is not at a satisfactory standard compared to the work and time, and if need to clock in overtime, then they are much more likely to leave the company. From a company perspective, if an employee is compensated well for their time and effort, which they can determine on an individual basis, then a straightforward recommendation for the company is to work with that employee to increase their income level, whether it be promotions or a bonus.

A second major determining factor for employee attrition is whether or not an employee has clocked in overtime. Using our entropy-based decision tree model, overtime is rooted as the top feature. After splitting the overtime variable into two variables, yes or no, and inputting it into the random forest model, using the Gini index, we see that both the lack of overtime —when overtime is zero— and the indication of overtime –or when overtime is one—are the second and third most important features out of the over thirty features.

Based on overtime, there are control methods or recommendations available to Human Resources. Firstly, if an employee is found to be both underpaid and working overtime, then this should trigger a flag for HR to evaluate that employee; they are essentially at high risk of leaving and it should be a concern. If this employee is determined to be of value to the company, then the superficial recommendation is that the company should investigate promotion possibilities and look into raising that employee's compensation package. They can then reevaluate the employee in six months and see if they continue to clock in overtime. If they continue to clock in overtime, or if there is a deeper set of reasons for why that employee requires overtime, then that employee needs to be brought in for an evaluation. If it turns out that they are clocking in overtime for an issue like lack of skill or because they cannot fulfill the requirements of the position, then more substantial recommendations will be needed, like training or an internal role transfer.

Another important attribute determining employee attrition is Job involvement. This does not directly relate to the Theory of organization equilibrium as does monthly income and overtime, yet it does coincide with research on employee attrition. The consensus is that if an employee is not more involved with the work that they do, then it is also more likely that they are also not satisfied with the job they have. This could be due to their level of satisfaction, the degree of available developmental opportunities, or other circumstances, like work-life balance, team dynamics, their manager, or even the distances that they need to travel to work. These reasons further stem into connections with their satisfaction, monthly income level, time, and effort.

Two additional attributes leading to notably higher attrition rates are the age of the employee and the extent at which they have worked with the company. We found that certain age groups, particularly those under the age of twenty-one, are significantly higher to leave the company than are others. When reviewing younger age groups, we surmise that these groups are either interns or temporary employees who have no real intent at staying with the company for an extended period. (review age groups and their relative salary; add this in – also review the age about the number of years worked). When looking at the number of years worked, we also find this to be of critical interest to human resources in evaluating attrition. We found that the first three years of an employee’s time with the company is the most critical. This time period should be carefully monitored by HR. If these employees begin clocking in overtime or put in requests for promotions and pay increases, then they are at a very high risk of leaving the company.

Dataset Overview

Dataset: IBM HR Analytics Employee Attrition & Performance

Data Dimensions	[1470 x 35]: 1470 Observations & 35 Features
Quality	High - No Missing Data
Target/label	Attrition
Source	https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
Value Key Categorical Attributes	Education: 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor' EnvironmentSatisfaction: 1 'Low' 2 'Medium' 3 'High' 4 'Very High' JobInvolvement: 1 'Low' 2 'Medium' 3 'High' 4 'Very High' JobSatisfaction: 1 'Low' 2 'Medium' 3 'High' 4 'Very High' PerformanceRating 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding' RelationshipSatisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High' WorkLifeBalance 1 'Bad' 2 'Good' 3 'Better' 4 'Best'

Resources

Machine Learning and Decision Trees

1. *Extension and Evaluation of ID3 – Decision Tree Algorithm*

    a. *https://pdfs.semanticscholar.org/bfae/5daa9ca4a6429b1ab837429f562d0a78df7d.pdf*

2. *Machine Learning Decision Trees Prof. Dr. Martin Riedmiller*

    a. *http://ml.informatik.uni-freiburg.de/former/_media/documents/teaching/ss11/ml/03_decisiontrees.printer.pdf*

3. *A comparative study of decision tree ID3 and C4.5*

    a. *https://saiconference.com/Downloads/SpecialIssueNo10/Paper_3-A_comparative_study_of_decision_tree_ID3_and_C4.5.pdf*

Research into HR and Employee Attrition

1. *Do You Really Know Why Employees Leave Your Company?*

    a. *https://hbr.org/2019/07/do-you-really-know-why-employees-leave-your-company*

2. *Why People Really Quit Their Jobs*

a. https://hbr.org/2018/01/why-people-really-quit-their-jobs

3. *Rational of Employee Attrition and Strategies to Employee Retention An Empirical Study*

a. https://www.researchgate.net/publication/329217411_Rational_of_Employee_Attrition_and_Strategies_to_Employee_Retention

4. *The Biggest Cost Of Doing Business: A Closer Look At Labor Costs*

a. https://www.paycor.com/resource-center/a-closer-look-at-labor-costs