PROJECT REPORT

IBM HR Analytics Employee Attrition & Performance

Abstract & keywords

Attrition is a critical issue and pretty high in the industry these days. It's the major problem which highlights in all the organizations. Though the term 'ATTRITION' is common, many would be at a loss to define what actually Attrition is, "Attrition is said to be the gradual reduction in the number of employees through retirement, resignation or death. It can also be said as Employee Turnover or Employee Defection" Whenever a well-trained and well-adapted employee leaves the organization, it creates a vacuum. So, the organization loses key skills, knowledge and business relationships. Modern managers and personnel administrators are greatly interested in reducing Attrition in the organization, in such a way that it will contribute to the maximum effectiveness, growth, and progress of the organization.

Retaining employees is a critical and ongoing effort. One of the biggest challenges in having managers in the place that understands it is their responsibility to create and sustain an environment that fosters retention. Staff requires reinforcement, direction and recognition to grow and remain satisfied in their positions. Managers must recognize this and understand that establishing such fundamentals demonstrates their objectives to support nature and motivate their employees.

The main objectives of this study are to know the reasons, why attrition occurs, to identify the factors which make employees dissatisfy, to know the satisfactory level of employees towards their job and working conditions and to find the areas where companiesare lagging behind.

Keywords: Attrition Analysis, Machine Learning, Oversampling

Certification of completion

I hereby certify that the project titled “IBM HR Analytics Employee Attrition &Performance” was undertakenand completed under my guidance and supervision by Arpit Rajput, Helly Bhalodia, Siddhanta Chhetri , Shriya Gupta students of the July A 2019 batch of the Post Graduate Program in Data Science & Engineering, Bangalore.

Mr.Srikar Muppidi

Date: 28th Nov 2019

Table of Contents

Abbreviations used

Abbreviation	Expansion
LR	Logistic Regression
DT	Decision Tree
AUC	Area Under the Curve
RF	Random Forest
LGBM	Light Gradient boosting method
Bag_DT	Bagging Decision Tree
Boost_DT	Boosting Decision Tree
FNR	False Negative Rate
FPR	False Positive Rate
XGB	Xtreme Gradient Boosting
SMOTE	Synthetic Minority Oversampling Technique
KNN	K-Nearest Neighbors

Executive summary

Background & need for study: Companies in India as well as in other countries face a formidable challenge of recruiting and retaining talents while at the same time having to manage talent loss through attrition be that due to industry downturns or through voluntary individual turnover. Losing talents and employees result in performance losses which can have long term negative effect on companies especially if the departing talent leaves gaps in its execution capability and human resource functioning which not only includes lost productivity but also possibly loss of work team harmony and social goodwill. The success of any organization depends largely on the workers, the employees are considered as the backbone of any company. The study was mainly undertaken to identify the level of employee's attitude, the dissatisfaction factors they face in the organization and for what reason they prefer to change their job. Once the levels of employee's attitude are identified, it would be possible for the management to take necessary action to reduce attrition level.

Scope & Objectives:

- To know the satisfactory level of employees towards their job and working conditions
- To identify the factors which make employees dissatisfy about company's policy and norms.
- To find the areas where companies is lagging behind
- To know the reasons, why attrition occurs in companies.
- To find the ways to reduce the attrition in companies.

Approach & methodology: The data is provided by IBM on Kaggle. After processing the dataset and cleaning the inconsistencies, the numerical and categorical features used in the attrition prediction model is generated. Various Classification algorithms are used to predict attrition based on set of independent variables like gender, workplace distance,stock level, salary hike etc. used. The predictive models are also used to identify the variables that strongly influence the attrition using variable importance and probabilistic approaches. The models are evaluated using relevant model performance measures to arrive at the most robust models for prediction.

Recommendations & actionable insights: The high-level recommendations for the project are developed by predicting attrition of an employee. These are then linked to the model findings to recommend actionable insights, which include speaking to employee on correct time to avoid attrition.

Chapter 1 - Project overview

Attrition is a problem that impacts all businesses, irrespective of geography, industry and size of the company. Employee attrition leads to significant costs for a business, including the cost of business disruption, hiring new staff and training new staff. As such, there is great business interest in understanding the drivers of and minimizing staff attrition.

In this context, the use of classification models to predict if an employee is likely to quit could greatly increase the HR's ability to intervene on time and remedy the situation to prevent attrition. While this model can be routinely run to identify employees, who are most likely to quit, the key driver of success would be the human element of reaching out the employee, understanding the current situation of the employee and taking action to remedy controllable factors that can prevent attrition of the employee.

Need for study

The success of any manufacturing organization depends largely on the workers, the employees are considered as the backbone of any company. The study was mainly undertaken to identify the level of employee's attitude, the dissatisfaction factors they face in the organization and for what reason they prefer to change their job. Once the levels of employee's attitude are identified, it would be possible for the management to take necessary action to reduce attrition level. Since they are considered as backbone of the company, their progression will lead to the success of the company for the long run. This study can be helpful in knowing, why the employees prefer to change their job and which factors make employee dissatisfy. Since the study is critical issue, it is needed by the originations in order to assess the overall interest and the feelings of the employees towards their nature of job and organization. This study can be helpful to the management to improve its core weaknesses by the suggestions and recommendations prescribed in the project. This study can serve as a basis for measuring the organization's overall performance in terms of employee satisfaction. The need of this study can be recognized when the result of the related study needs suggestions and recommendations to the similar situation.

Current baseline & Business mission

The average employee attrition in Analytics & Data Science Industry in India currently stands at 24.4% annually.Also, analytics employees tend to stay in their organization for an average of 3.7 Years.

Problem statement & project scope

The aim of the present report is to study various factors like salary, growth opportunities, facilities, policies and procedures, recognition, appreciation, suggestions by which it helps to know the Attrition level in the organizations and factors relating to retain them. This study also helps to find out which employee are likely to leave organization.

Data sources

(Data source: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.)

This data set presents an employee survey from IBM, indicating if there is attrition or not. The data set contains approximately 1500 entries. Given the limited size of the data set, the model should only be expected to provide modest improvement in identification of attrition vs a random allocation of probability of attrition.

While some level of attrition in a company is inevitable, minimizing it and being prepared for the cases that cannot be helped will significantly help improve the operations of most businesses. As a future development, with a sufficiently large data set, it would be used to run a segmentation on employees, to develop certain “at risk” categories of employees. This could generate new insights for the business on what drives attrition, insights that cannot be generated by merely informational interviews with employees.

IBM has gathered information on employee satisfaction, income, seniority and some demographics. It includes the data of 1470 employees. To use a matrix structure, we changed the model to reflect the following data.

Dataset Description

During website session, browsing information about visited pages is collected and features are extracted as follows

Table 1 – Numerical features used in the user attrition analysis model

Feature Name	Feature Description	Min Value	Max Value	StdDev
Age	Age of employee	18	60	9.13
DailyRate	It is the billing cost for an individual's services for a single day	102	1499	403.50
DistanceFromHome	It is the distance between company and home of the employee	1	29	8.10
Education	Education qualification of the employees of company	1	5	1.02
EmployeeCount	Count of employee	1	1	0.0
EmployeeNumber	It is a unique number that has been assigned to each current and formeremployee	1	2068	602.02
EnvironmentSatisfaction	It is all about an individual's feelings about the work environment andorganization culture.	1	4	1.09
HourlyRate	The amount of money that is paid to an employee for every hour worked	30	100	20.32
JobInvolvement	Job involvement refers to the degree to which a job is central to a person's identity.	1	4	0.71
JobLevel	Job levels are categories of authority in an organization.	1	5	1.10
JobSatisfaction	Job satisfaction happens when an employee feels he or she is having job stability.	1	4	1.10
MonthlyIncome	Gross monthly income is the amount of income an employee earn in one month.	1009	19999	4707.95
MonthlyRate	If a monthly rate is set, employees should be paid in exchange for normal hours of work of a full-timeworker.	2094	26999	7117.78
NumCompaniesWorked	Number of other companies the employee previously worked for	0	9	2.49
PercentSalaryHike	The amount a salary is increased of an employee in percentage	11	25	3.65
PerformanceRating	Rating means gauging and comparing the performance.	3	4	0.36
RelationshipSatisfaction	It is the rate of satisfaction between employer–employee relationship.	1	4	1.08

StandardHours	Standard Hours employee is working	80	80	0.0
StockOptionLevel	Employee stock options.	0	3	0.85
TotalWorkingYears	Total number of years employee worked	0	40	7.78
TrainingTimesLastYear	No of months the employee is trained by the company.	0	6	1.28
WorkLifeBalance	Work-life balance refers to the level of prioritisation between personal and professionalactivities in an individual's life.	1	4	0.70
YearsAtCompany	Total Number of Years at the Company	0	40	6.12
YearsInCurrentRole	Number of years employee worked in current role	0	18	3.62
YearsSinceLastPromotion	Number of years of an employee since last promotion	0	15	3.22
YearsWithCurrManager	Number of years employee worked with current manager	0	17	3.56

Table 1 Shows the numerical features along with their statistical parameters.

Table 2 – Categorical Features used in the User Attrition Analysis Model

Feature Name	Feature Description	Number of Categorical Values
Attrition	Attrition in business describes a gradual but deliberate reduction in staff numbers that occurs as employees retire or resign, [NOTE: Target Variable] (0=no, 1=yes)	2
BusinessTravel	Business travel is travel undertaken for work or business purposes, as opposed to other types of travel (1=No Travel, 2=Travel Frequently, 3=Travel Rarely)	3
Department	Consists three departments that contribute to the company's overall mission. (1=HR, 2=R&D, 3=Sales)	3
EducationField	Education field of the employees(1=HR, 2=Life Sciences, 3=Marketing, 4=Medical Sciences, 5=others, 6= Technical)	6
Gender	Gender of the employee (1=Female, 2=Male)	2
JobRole	These refer to the specific activities or work that the employee will perform. (1=HC Rep, 2=HR, 3=Lab Technician, 4=manager, 5= Managing Director, 6= Research Director, 7= Research Scientist, 8=sales Executive, 9= Sales Representative)	9
MaritalStatus	Marital Status of the employee (1=divorced, 2=married, 3=single)	3
Over18	(1=Yes, 2=No)	2
Overtime	(1=No, 2=Yes)	2

Table 2 Shows the categorial features along with their number of categories.

Data preparation & clean up

The source dataset received has been prepared to ensure that the fields are cleaned up, the values are suitable for model building and the variable names are self-explanatory. The broad approach for data preparation can be outlined as:

Table 4 – Data pre-processing steps

Label Encoding	Outlier Treatment	Standardization	Oversampling
Categorical Variable BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, OverTime are converted into Dummy variable	Box plot is drawn for Independent features against Target variable and outlier had been detected.	Standard Scalar function from Scikit learn library since the numerical variable are of different scale in order to obtain better performance.	Since our Dataset is highly imbalanced, we used SMOTE oversampling technique in order to tackle class imbalance.
Attrition Feature is converted into binary value 0's and 1's	Since the outliers are legitimate, we have decided to retain them in data		

Statistical tools & techniques

The dependent variable is whether the employee will leave the organization or not based on various factors

The model building exercise has also considered cross validation and tuning techniques to ensure that the models built perform well when used for prediction.

The classification algorithms used for Commercial intent prediction include

- Logisticregression
- DecisionTree
- RandomForest
- GradientBoosting

Model performance measures used for evaluating models

The various models built, must be evaluated based on certain model performance measures to identify the most robust models. The choice of the right model performance measures is highly critical since the dataset is a highly imbalanced dataset and the attrition rate is 16%. Model accuracy alone may not be enough to evaluate a model. Hence the following model performance measures have been used to evaluate the models, based on the confusion matrix built for the predictions on the training and test datasets:

	Negative (Predicted)	**Positive (Predicted)**
Negative (Observed)	True Negative (TN)	False positive (FP)
Positive (Observed)	False negative (FN)	True positive (TP)

Accuracy

Accuracy is the number of correct predictions made by the model by the total number of records. The best accuracy is 100% indicating that all the predictions are correct.
Considering the response rate (conversion rate) of our dataset which is ~16%, accuracy is not a valid measure of model performance. Even if all the records are predicted as 0, the model will still have an accuracy of 84%. Hence other model performance measures need to be evaluated.

Sensitivity or recall

Sensitivity (Recall or True positive rate) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall or true positive rate (TPR).
For our dataset, it gives the ratio of actual customers who generated revenue by the total number of customers predicted who will generate the revenue.

Specificity

Specificity (true negative rate) is calculated as the number of correct negative predictions divided by the total number of negatives.
For our dataset, specificity gives the ratio of actual customers who will not generate revenue by the number of customers who are predicted who will not generate revenue.

Precision

Precision (Positive predictive value) is calculated as the number of correct positive predictions divided by the total number of positive predictions.
Precision tells us, what proportion of customers who generated revenue as customers actually generated revenue. If precision is low, it implies that the model has lot of false positives.

F1-Score

F1 is an overall measure of a model's accuracy that combines precision and recall A good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0.

ROC chart & Area under the curve (AUC)

ROC chart is a plot of 1-specificity in the X axis and sensitivity in the Y axis. Area under the ROC curve is a measure of model performance. The AUC of a random classifier is 50% and that of a perfect classifier is 100%. For practical situations, an AUC of over 70% is desirable.

Level of significance

For all the hypothesis tests in the project, the level of significance is assumed as 5% unless specified otherwise.

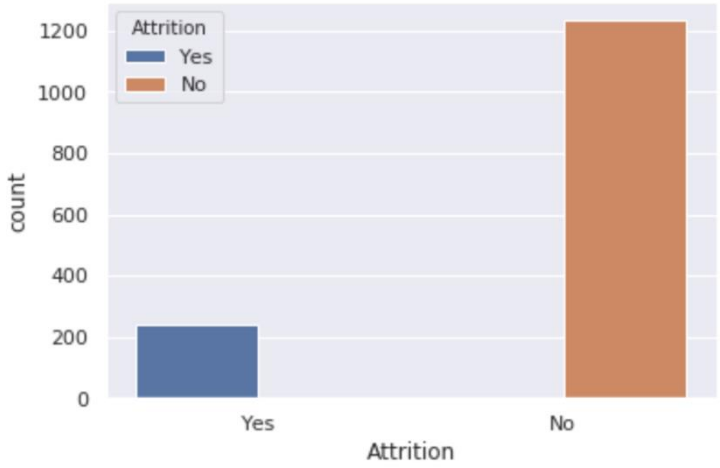
Chapter 2 - Exploratory data analysis

The purpose of exploratory data analysis is two-fold:

- To understand the data in terms of attrition information across various independentvariables
- Get insights on variousfeatures.

Understand data distribution

OverallAttrition

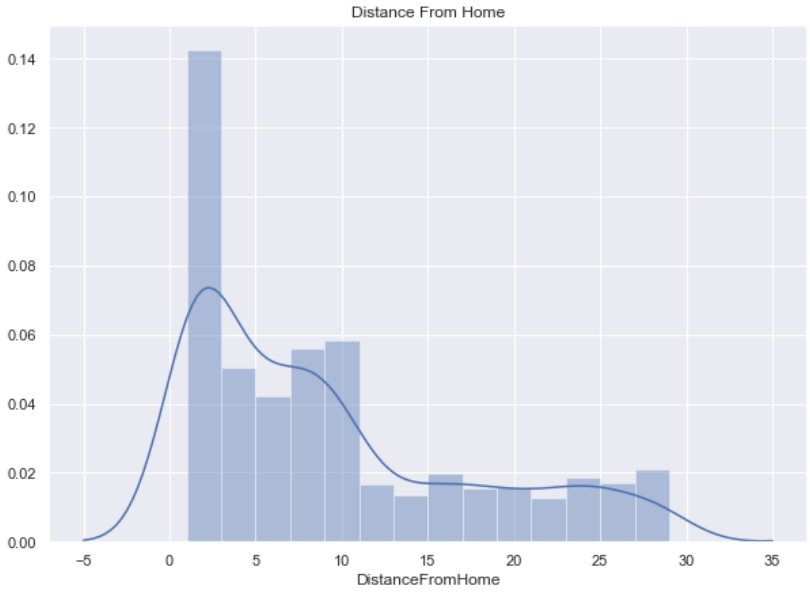


The attrition rate is = 237/1470 = 16.1 %.

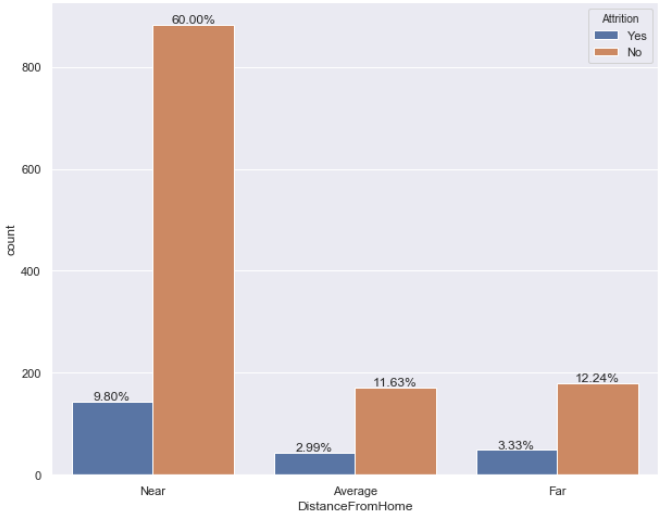
- This indicates that the data set is an imbalanced dataset where the number of observations belonging to class 1 (No) is significantly higher than those belonging to class 0 (True)
- The conventional accuracy of the predictive models is not a relevant measure of model performance because machine learning algorithms are usually designed to improve accuracy by reducing the error. Thus, they do not take into account the class distribution / proportion or balance of classes.
- Hence, we will consider other model performance measures to evaluate a model, keeping in mind the class imbalance problem.

Exploratory Data Analysis

Distribution of Distance from Home

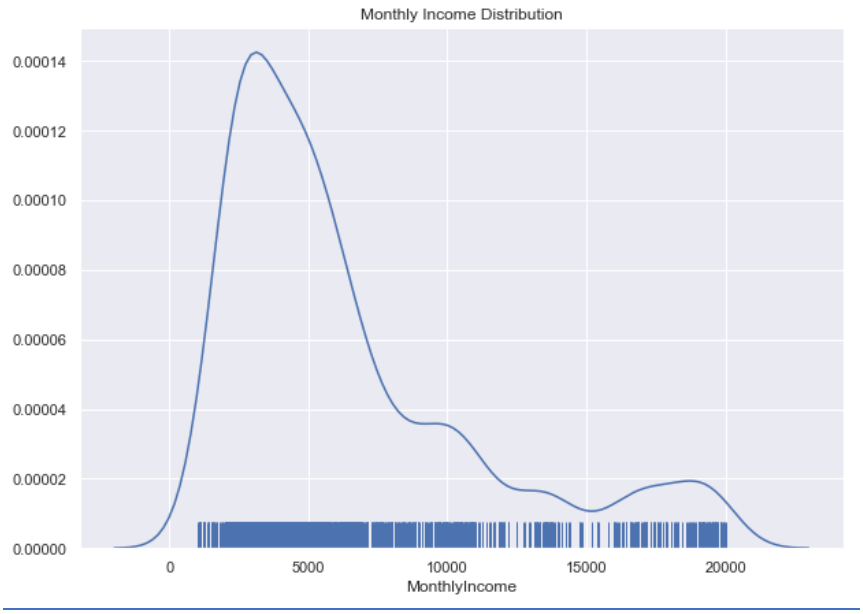


- Our distribution of Distance from Home feature is positively skewed implying most of the people lives nearby.
- Doing Chi2_contingency statistical test shows that the feature is significant and dependent.

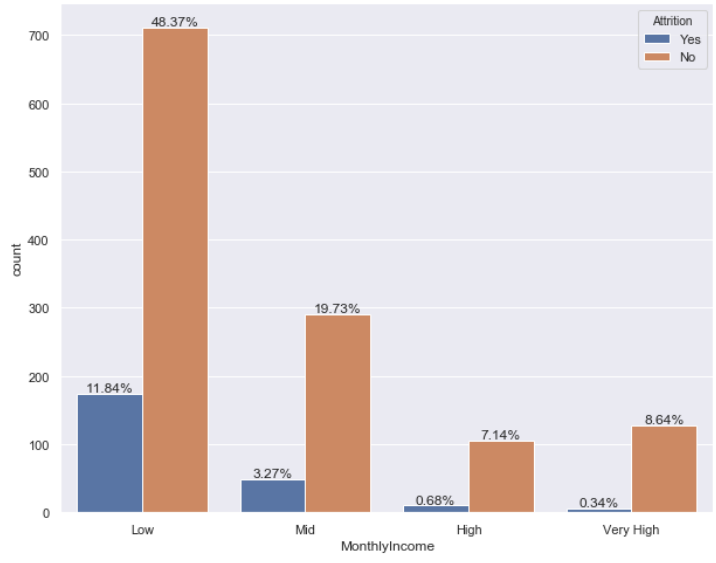


	Attrition	
DistFromHome	**Yes**	**No**
Near	9.80%	60.00%
Average	2.99%	11.63%
Far	3.33%	12.24%

Distribution of Monthly Income

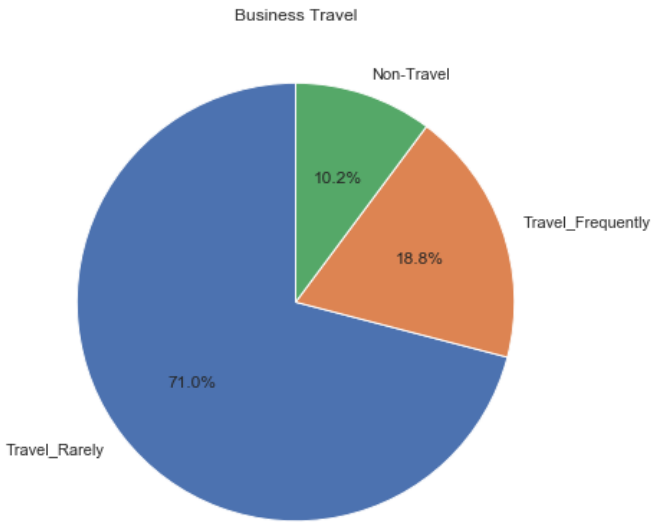


- Our distribution of Monthly Income is positively skewed, implying there are less people in high monthly income bracket.
- Converting continuous to discrete and doing statistical test implies that this feature is not significant and independent.
- The highest Attrition rate is in Low Income Bracket, while the lowest in High and Very High Income Barcket.



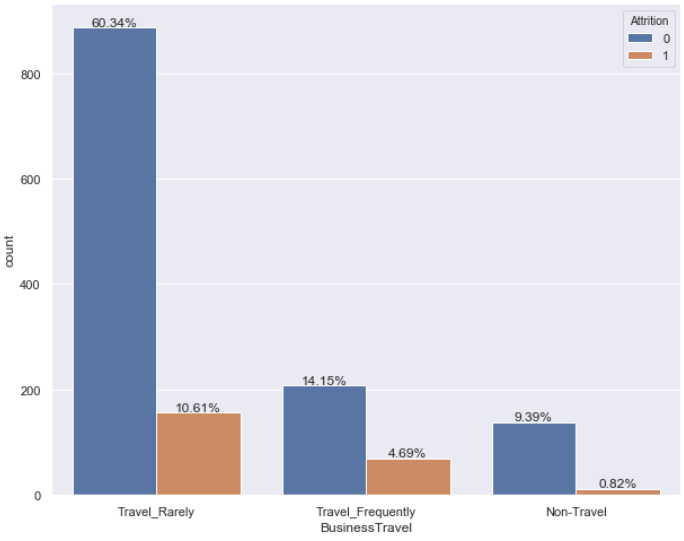
	Attrition	
MonthlyIncome	**Yes**	**No**
Low	11.84%	48.37%
Mid	3.27%	19.73%
High	0.68%	7.14%
Very High	0.34%	8.64%

Distribution of Business Travel



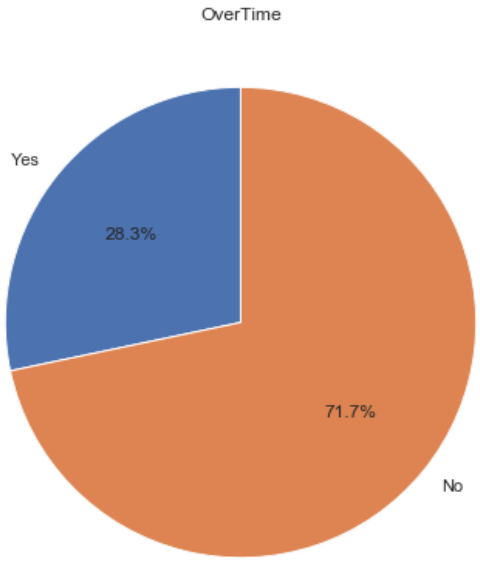
BusinessTravel	Distribution
Travel Rarely	71.0%
Travel Frequently	18.8%
Non Travel	10.2%

- There are high number of employees who travel rarely, followed by those employee who travel frequently and non - travellers.
- The highest number of Attrition rate is of the employee who travels rarely, while the lowest being the non - traveller one.



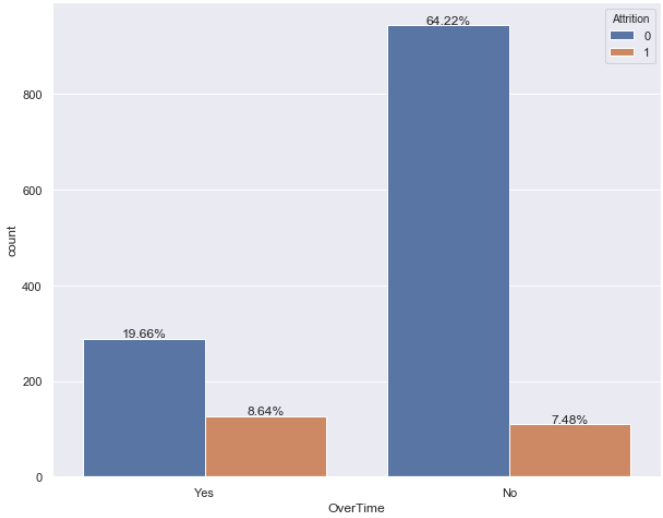
	Attrition	
BusinessTravel	Yes	No
Travel Rarely	10.61%	60.34%
Travel Frequently	4.69%	14.15%
Non Travel	0.82%	9.39%

Distribution of Over Time employee



	Attrition	
OverTime	**Yes**	**No**
	71.70%	28.30%

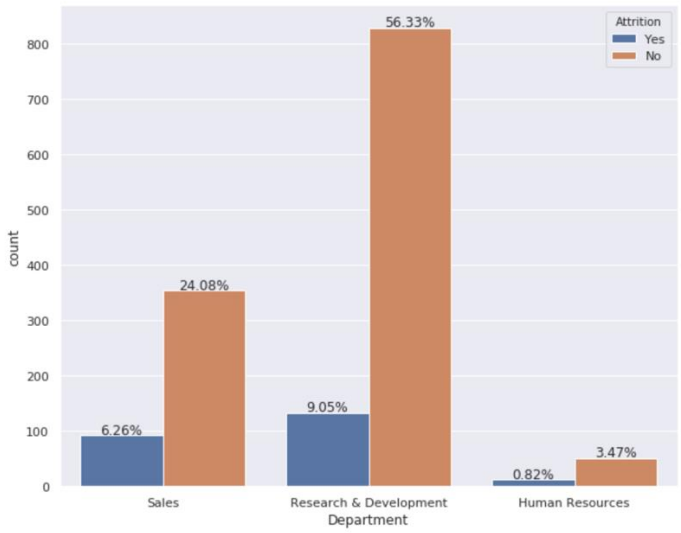
- Employee who does Overtime (28.30%) is less than those who do not (71.70%).
- Attrition Rate of Overtime doer (8.64%) is higher than their counterpart (7.48%)



	Attrition	
OverTime	**Yes**	**No**
Yes	8.64%	19.66%
No	7.48%	64.22%

Other Exploratory Data Analysis

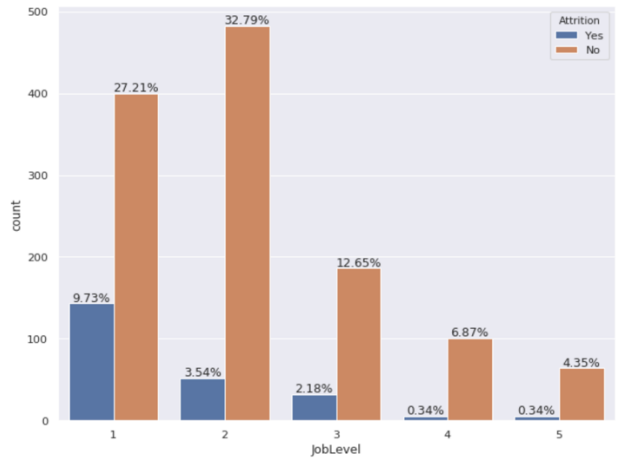
Attrition by Department



Department	Yes	No
Sales	6.26%	24.08%
R & D	9.05%	56.33%
HR	0.82%	3.47%

- Majority of the employees are from R&D department and sales.
- R&D depart has the highest attrition rate- Could be a consequence of the higher number of people from R&D.
- Overall Attrition is more in sales as it has a smaller number of employees and more attrition as compared with R&D.

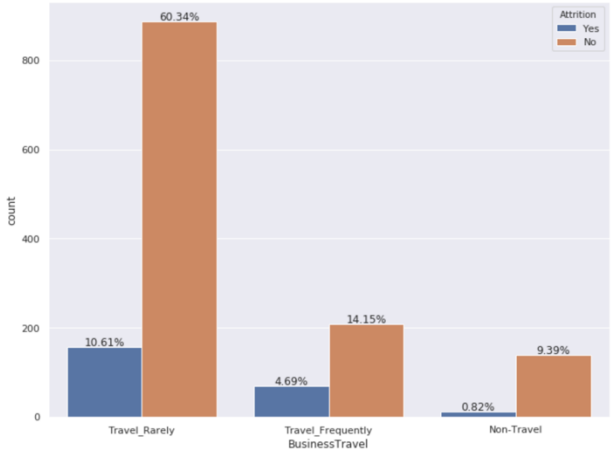
Attrition by Job Level



Job Level	Yes	No
1	9.73%	27.21%
2	3.54%	32.79%
3	2.18%	12.65%
4	0.34%	6.87%
5	0.34%	4.35%

- Majority of the employees are in lower and intermediate level.
- Most Attrition is from lower level, followed by intermediate level.
- Figure shows that juniors are leaving the company most – addressing problems faced by lower level at the company could help reduce attrition.

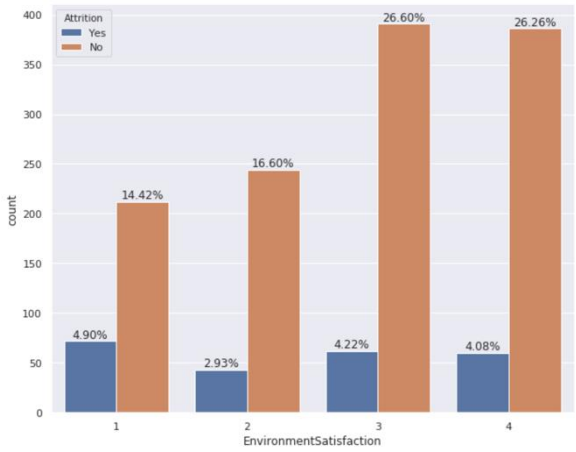
Attrition by Travel Frequency



Frequency	Yes	No
Rarely	10.61%	60.34%
Frequently	4.69%	14.15%
No Travel	0.82%	9.39%

- Majority of the employees travel rarely.
- Also, it has high attrition rate.
- On the other hand, who do not travel are very rare to leave company.
- So, it could be on site facilities are not good. Improving them could reduce the attrition.

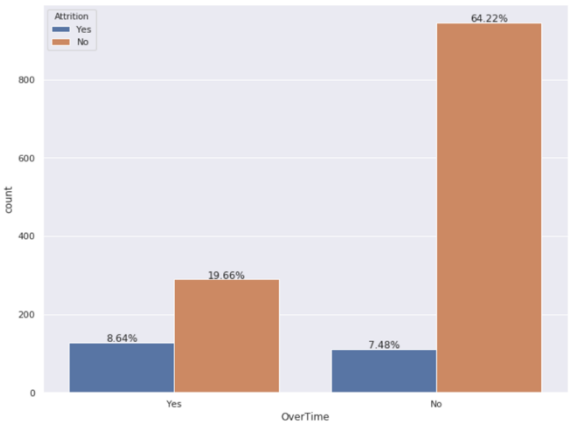
Attrition by Environment Satisfaction



Env. Satisfaction	Yes	No
1	4.90%	14.42%
2	2.93%	16.60%
3	4.22%	26.60%
4	4.08%	26.60%

- By looking into the bar graph, it is visual that environment satisfaction does not play major role in attrition.
- Though low satisfaction has little more attrition but not major difference.

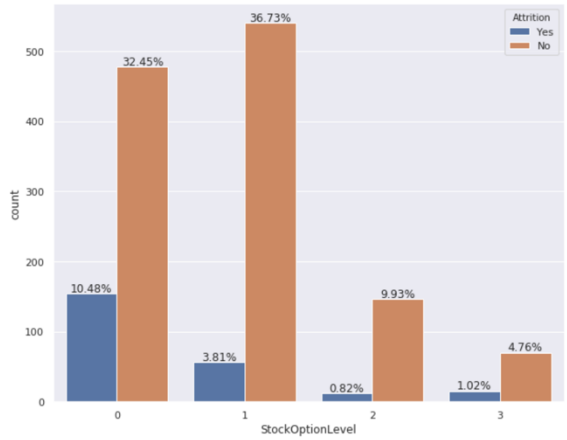
Attrition by Over Time



Over Time	Yes	No
Yes	8.64%	19.66%
No	7.48%	64.22%

- People who do over time are likely to leave company more.
- It could be due to over workload.
- Discussion on the regular basis about the workload, may solve this problem.

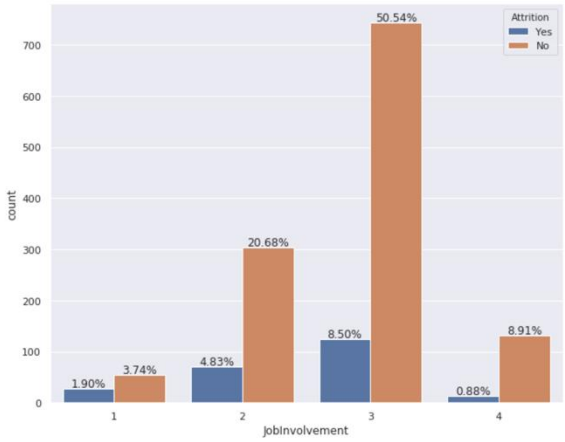
Attrition by Stock Option Level



Stock Option Level	Yes	No
0	10.48%	32.45%
1	3.81%	36.73%
2	0.82%	9.93%
3	1.02%	4.76%

- High stock option is inversely proportional to the attrition rate.
- Employees with less stock options leave the company more often.
- So, providing more stock options could reduce the attrition rate.

Attrition by Job Involvement



Job Involvement	Yes	No
1	1.90%	3.74%
2	4.83%	20.68%
3	8.50%	50.54%
4	0.88%	8.91%

- Graph show that most employees are highly engaged with job.
- It is obvious from the graph that employee with high involvement tend to leave more.
- May be reducing work load little bit and giving them some free time to think on some innovation activities will reduce the attrition rate.

Correlation Matrix for Numerical Features

From the correlation plots, we can see that quite a lot of our columns seem to be poorly correlated with one another. Generally, when making a predictive model, it would be preferable to train a model with features that are not too correlated with one another so that we do not need to deal with redundant features. In the case that we have quite a lot of correlated features one could perhaps apply a technique such as Principal Component Analysis (PCA) to reduce the feature space.

Chapter 3 - Feature Selection & Model Building

Feature selection is the process of selecting a subset of relevant attributes to be used in making the model in machine learning. Effective feature selection eliminates redundant variables and keeps only the best subset of predictors in the model which also gives shorter training times. Besides this, it avoids the curse of dimensionality and enhance generalization by reducing overfitting.

In this project, feature selection techniques are applied to improve the classification performance and/or scalability of the system. Thus, we aim to investigate if better or similar classification performance can be achieved with a smaller number of features. An alternative of feature selection is the use a feature extraction technique such as select K-Best, Backward elimination, Principal Component Analysis for dimensionality reduction. However, in this case, the features in the reduced space will be the linear combinations of 35 attributes, which brings the need of tracking all features. Therefore, it has been deemed appropriate to apply feature selection instead of feature extraction within the scope of this research. For feature ranking, instead of wrapper algorithms that require a learning algorithm to be used and consequently can result in reduced feature sets specific to that classifier, filter-based algorithms are tested in which no classification algorithm is used. Correlation Attribute Evaluation, Information Gain Attribute Evaluation and Minimum Redundancy Maximum Relevance Filters were used in our experiments. Thus, maximum classification accuracy is aimed to be obtained with minimal subset offeatures.

Classification Results:

One of the purposes of this project is to get the analyses results of the measuring whether a employee is going to leave organization. The dataset is fed to Logistic Regression, Decision tree, Random Forest, Naive bayes and Gradient Boosting classifiers. The Accuracy, ROC Score and F1-Score are presented for each classifier.

Results on class imbalanced dataset:

Tables below show the results obtained with Logistic Regression, Decision Tree, Random forest, gradient boosting and Decision Tree. The results show that Logistic Regression gives the highest F1 Score on test set. However, a class imbalance problem arises since the number of negative class instances in the data set is much higher than that of the positive class instances, and the imbalanced success rates on positive (TPR) and negative (TNR) samples show that the classifiers tend to label the test samples as the majority class. This class imbalance problem is a natural situation for the problem.

Algorithm	Accuracy	F1-score
LR	0.870748	[0.925,0.521]
DT	0.832517	[0.906,0.177]
RF	0.834136	[0.908,0.120]
NB	0.770340	[0.855,0.514]
GB	0.842143	[0.908,0.406]

Results obtained with Oversampling:

The results presented in this section show that the classifiers tend to minimize their errors on majority class samples, which leads to an imbalance between the accuracy rates of the positive and negative classes. However, in a real-time attrition analysis model, correctly identifying attrition, which are represented with positive class in our dataset, is as important as identifying negative class samples. Therefore, a balanced classifier is needed to increase the conversion rates in an e-commerce website. To deal with class imbalance problem, we use oversampling method, in which a uniform distribution over the classes is aimed to be achieved by adding more of the minority (positive class in our dataset) class instances. Since this dataset is created by selecting multiple instances of the minority class more than once, first oversampling the dataset and then dividing it into training and test sets may lead to biased results due to the possibility that the same minority class instance may be used both for training and test. For this reason, in our study, 30 percentage of the data set consisting of 494 samples is first left out for testing and the oversampling method is applied to the remaining 70 percentage of the samples.

The results obtained on the balanced dataset are shown in the table below. Since the number of samples belonging to positive and negative classes is equalized with oversampling, both accuracy and F1-score metrics can be used to evaluate the results.

Algorithm	Accuracy	F1-score
LR	0.787547	[0.775,0.801]
DT	0.748911	[0.738,0.758]
RF	0.834348	[0.833,0.834]
NB	0.704438	[0.647,0.745]
GB	0.925955	[0.923,0.926]

Insights inferred through odds-Ratio using Logistic Regression:

Using the coefficients of the X variables in the new model, the odds ratio and probability have been calculated as follows:

Odds ratio = exp(coef(LR model))

Probability = Odds/(1+Odds)

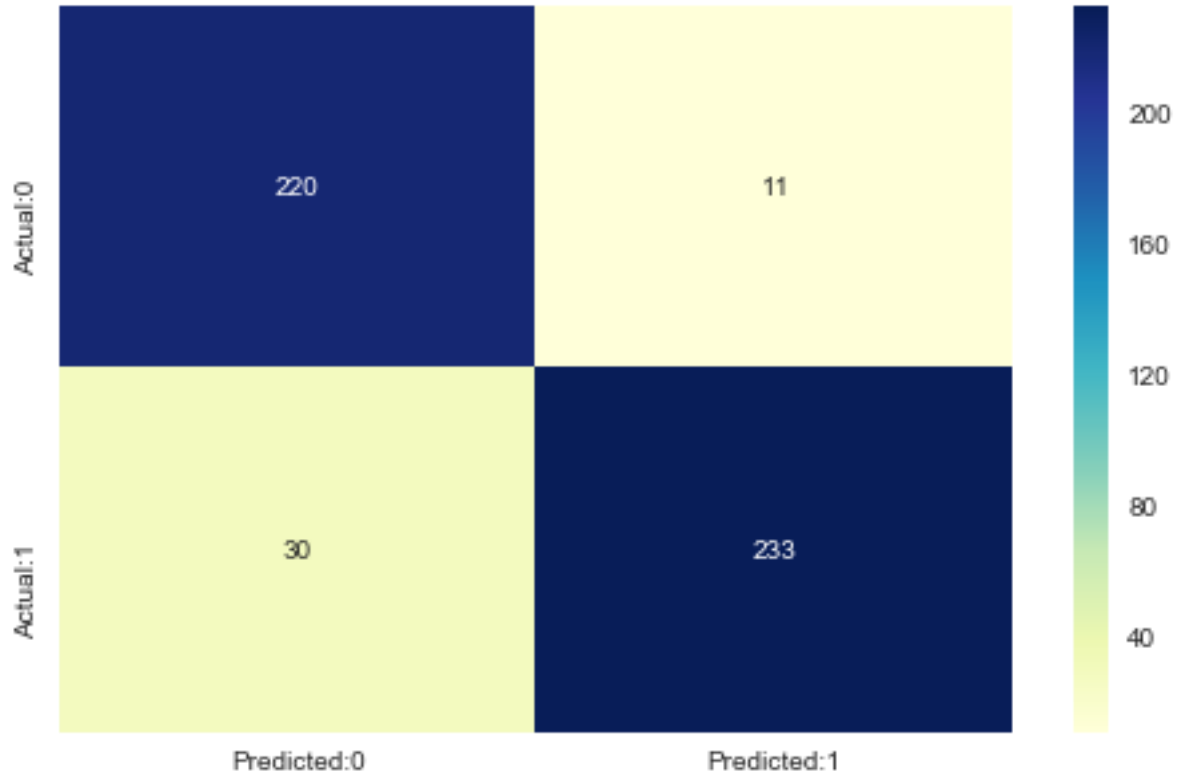
The Odds ratio and probability of some of the variables, based on their practical significance have been listed below:

Predictor Variable	Co-eff	Odds Ratio	Probability
Age	-0.0298	0.970596	0.492539
DistanceFromHome	0.0437	1.044617	0.510911
EnvironmentSatisfaction	-0.4323	0.649035	0.393585
JobInvolvement	-0.5264	0.590715	0.371352
JobSatisfaction	-0.4233	0.654898	0.395733
NumCompaniesWorked	0.1886	1.207557	0.547010
RelationshipSatisfaction	-0.2505	0.778447	0.437712
TotalWorkingYears	-0.0795	0.923614	0.480145
TrainingTimesLastYear	-0.2011	0.817805	0.449886
WorkLifeBalance	-0.3886	0.677973	0.404043
YearsAtCompany	0.0829	1.086398	0.520705
YearsInCurrentRole	-0.1321	0.876291	0.467034
YearsSinceLastPromotion	0.1773	1.194022	0.544216
YearsWithCurrManager	-0.1313	0.876998	0.467235
BusinessTravel_Travel_Frequently	1.9251	6.855935	0.872708
BusinessTravel_Travel_Rarely	1.0493	2.855793	0.740650
EducationField_Technical Degree	0.8928	2.441908	0.709463
Gender_Male	0.3852	1.469922	0.595129
JobRole_Human Resources	1.2936	3.646033	0.784763
JobRole_Laboratory Technician	1.1079	3.027917	0.751733
JobRole_Sales Executive	0.8628	2.369864	0.703252
JobRole_Sales Representative	1.6970	5.457586	0.845143
MaritalStatus_Single	1.1604	3.191338	0.761413
OverTime_No	1.8319	6.245541	0.861984
OverTime_Yes	3.8075	45.039862	0.978280

- If the travel frequency is increased attrition will be reduced to 87%
- If over time is continued attrition will be impacted by probability of 97%

Chapter 4 - Conclusions

This project implemented predictive analyses on employee attrition by effective feature selection using a hybrid model of ensemble methods based on the CRISP-DM business model for IBM which is a multinational technology company. Different types of ensemble methods like boosting and bagging were tested for classification, however boosting algorithms performed the best but G boosting gave the best evaluation scores with getting up to 92% accuracy. Hence Gradient Boosting with oversampling can be taken as final model.



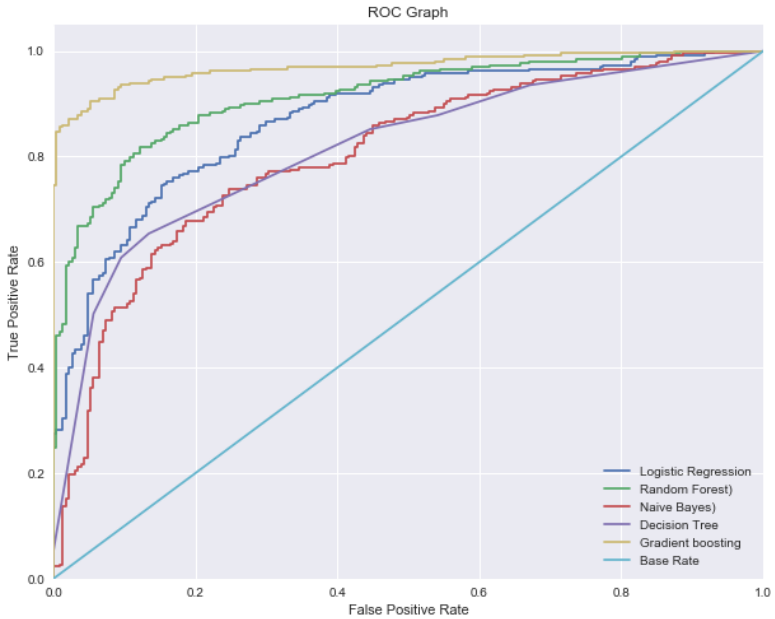
Following shows the confusion matrix for the model selected i.e, Gradient Boosting with over sampling.

The acuuracy of the model = TP+TN / (TP+TN+FP+FN) = 0.917004048582996
The Miss-classification = 1-Accuracy = 0.082995951417004
Sensitivity or True Positive Rate = TP / (TP+FN) = 0.8859315589353612
Specificity or True Negative Rate = TN / (TN+FP) = 0.9523809523809523
Positive Predictive value = TP / (TP+FP) = 0.9549180327868853
Negative predictive Value = TN / (TN+FN) = 0.88
Positive Likelihood Ratio = Sensitivity / (1-Specificity) = 18.604562737642567
Negative likelihood Ratio = (1-Sensitivity) / Specificity = 0.1197718631178707

Variable Importance Plot for Tree Based Algorithms

Gradient Boosti ng:

ROC Curve Comparison:



From the above figure it is clear that Gradient Boosting has the maximum ROC compared to other techniques.

In this first case study, it showed that area under the curve is higher in GB with higher accuracy and second case study showed that employees who are working at entry-level have higher chance to leave the organisation. Similarly, third case study explains that stock option level is an important variable

factor in determining employee attrition.

These case studies further evaluated the research question by showing the result that ensemble method with effective feature selection are effective in predicting employee attrition contributing to Human resource management sector which can be seen by visualizations and accuracies by different models thus managers should look into the top needs of the employee by motivating entry level employees, giving more stock option level, increasing job satisfaction, relationship satisfaction and avoiding overtime by employees.

However, this project has some limitations. This research is limited to a small dataset which lacks to train the model well that might give low results and getting employees data from an organisation is confidential hence this research is limited to IBM dataset which is the only available dataset online. The second drawback is with the model is limited to only supervised machine learning that requires a lot of computation time, sometimes decision boundary might be over trained that and user input is required every time when new features have to be added. This project can be extended in future as it has a lot of potentials to improve by applying deep learning techniques with a well-designed network of sufficient hidden layers on big data set which can cover up the limitations of this project. There can be time series and trend analysis which might improve the prediction performance if the data is in date format.

Chapter 5 – Recommendations and actionable Insights

Predictive analytics comes in picture when planning a mitigating action plan towards addressing the attrition rates; this in turn would help in working towards developing tools and techniques to reduce the impact of attrition as well.

Reasons why predictive analysis should be used for attrition and loyalty analysis for employees:

- *Gives a multidimensional analysis by drawing relationship within data sets:* Using predictive algorithms lead to a significant improvement in studying employee turnover and beats the degree of insights of any traditional, descriptive approach by 20 to 50%.
- *Can be used to analyze large sets of historical data to find patterns in the behavior:* Using averages and large aggregations instead of diving deep in the historical and individual patterns of turnover.
- *Provides with 'what-if' scenarios to help in decision making*: Develop models which allow them to try out different alternatives and to vary the constraints and the assumptions in order to see how the results would change.
- *Future estimation based on existing data can be done:* Assessing impactful problems such as key employee turnover or absenteeism can benefit enormously of predictive analytics compared to the use of historical reporting.

Now that We had a machine learning model that can predict whether or not a member of staff may leave and we had an idea of what contributes to job satisfaction, we had what we needed to build a system that creates recommendations for improving job satisfaction for employees who may be leaving soon. We thought it useful to create a "retention profile" which summarizes the traits of the members of staff who did not leave the organization. Doing so would allow me to compare the profile of a staff member who is likely to leave with this "retention profile" and easily generate insights that are actionable.

Prescriptive Analytics can be useful for bolstering the decision-making power of staff who may be new to managerial positions and may be acting or being trained for such positions. We suppose that in some cases it can even be used to for evaluating potential candidates for such positions.

We personally feel that for this type of scenario, Prescriptive Analytics can be useful for getting to the root cause of employee dissatisfaction and finding ways to strengthen the relationship between staff and management. The best way to see if this system is truly capable of delivering value, would be to try it out in real life.

Chapter 6 - References and Bibliography

- Dutta, D., Gaspar, B., Challenger, J. and Arora, D. (2010). Determining employee characteristics using predictive analytics. US Patent App. 12/814,756.

- Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms, algorithms.

Chapter 7 – Appendix

Detailed data dictionary

Feature Name	Feature Description
Age	Age of the employee, Numerical, int, Continuous
Attrition	Attrition in business describes a gradual but deliberate reduction in staff numbers that occurs as employees retire or resign, [NOTE Target Variable], object, Discrete
Business Travel	Business travel is travel undertaken for work or business purposes,as opposed to other types of travel, Categorical, object
Daily Rate	It is the billing cost for an individual's services for a single day.It is sometimes called a per diem, Numerical, int
Department	Consists three departments that contribute to the company's overall mission. Department includes Sales, Research and Development, Human Resource., Categorical, object
DistanceFromHome	It is the distance between company and home of the employee. Not clearly stated what unit of distance is mentioned, speculationkm/mile, Numerical, int
Education	Education qualification of the employees of company1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor', Categorical, int
Education Field	Education field of the employeesFields consists of Life Sciences, Medical, Marketing,Technical Degree, Human Resources, Other (consists of all remaining field present in company),Categorical, object
Employee Number	It is a unique number that has been assigned to each current and formeremployee. As new employees are hired or appointed they will be assigned the next available number. Numerical, int
Environment Satisfaction	It is all about an individual's feelings about the work environment andorganization culture.[NOTEThis does not include work, job security and so on as it belongs to Job Satisfaction]
EnvironmentSatisfaction	1 'Low' 2 'Medium' 3 'High' 4 'Very High',Categorical, int
Gender	Gender of the employee (Female, Male), Categorical, object
Hourly Rate	The amount of money that is paid to an employee for every hour workedNumerical, int
Job Involvement	Job involvement refers to the degree to which a job is central to a person's identity.From employee's perspective job involvement constitutes a key to motivation, performance,personal growth, and satisfaction in the workplace.
JobInvolvement	1 'Low' 2 'Medium' 3 'High' 4 'Very High', Categorical, int
Job Level	Job levels are categories of authority in an organization. Each level is typically associated with a salary range and a series of job titles.Higher the level more salary and authority. Categorical, int, ordinal
Job Role	These refer to the specific activities or work that the employee will

	perform.A "job title" is a convenient name for a role. Here Job Title is mentioned.Categorical, Object
Job Satisfaction	Job satisfaction happens when an employee feels he or she is having job stability,job security, career growth and a comfortable work life balance.
JobSatisfaction	1 'Low' 2 'Medium' 3 'High' 4 'Very High', categorical, int, ordinal
Marital Status	Marital Status of the employee, Category, object
Monthly Income	Gross monthly income is the amount of income an employee earn in one month, before taxes or deductions are taken out.Gross Monthly Income = Annual Salary / 12 (Numerical, int)
Monthly Rate	If a monthly rate is set, employees should be paid in exchange for normal hours of work of a full-time worker. This excludes overtime hours per month or a recurring bonus/commissionGross monthly Rate = (Hourly Pay) * (Hours/Week) *52 / 12
Num Companies Worked	Number of other companies the employee previously worked for, Categorical, int
OverTime	Whether the employee did overtime or not, Categorical, object
Percent Salary Hike	the amount a salary is increased of an employee in percentage, numerical, int
Performance Rating	Rating means gauging and comparing the performance or pace rate of a worker against a standard performance level set by the study analyst.Performance Rating = – Observed Performance/Normal Performance x 100, Categorical, int, ordinal
Relationship Satisfaction	It is the rate of satisfaction between employer–employee relationship. (Categorical, int, ordinal)
Standard Hours	Standard hours which is enforced by the company
Stock Option Level	Employee stock options, also known as ESOs, are stock options in the company's stock granted by an employerto certain employees. (Categorical, int, ordinal)
Total Working Years	Total number of years employee worked with current or other employer, numerical, int
Training Times last year	No of months the employee is trained by the company.[Unit of Time is not mentioned, speculationMonth] (Categorical, int)
Work Life Balance	Work-life balance refers to the level of prioritization between personal and professionalactivities in an individual’s life. It helps company to calculate job satisfaction level of employeeWorkLifeBalance 1 'Bad' 2 'Good' 3 'Better' 4 'Best' (categorical, int, ordinal)
Years at Company	Number of years employee worked with current company. numerical, int
Years In Current Role	Number of years employee worked in current role, numerical, int
Years Since Last Promotion	Number of years of an employee since last promotion, numerical, int
YearsWithCurrManager	Number of years employee worked with current manager (numerical int)