A
Project Report
On
"Prediction of Employee Attrition using Data Mining Techniques"

Prepared by

Kanksha Masrani (14IT058)

Under the guidance of

Prof. Jalpesh Vasa

A Report Submitted to

Charotar University of Science and Technology

For Partial Fulfillment of the Requirements for the

Degree of Bachelor of Technology

In Information Technology

(8th Semester Software Project Major-IT407)

Submitted at

DEPARTMENT OF INFORMATION TECHNOLOGY

Chandubhai S. Patel Institute of Technology

At: Changa, Dist: Anand – 388421

April 2018

CANDIDATE'S DECLARATION

I hereby declare that the project entitled **"Prediction of Employee Attrition using Data Mining Techniques"** is my own work conducted under the guidance of **Prof. Jalpesh Vasa.**

I further declare that to the best of my knowledge, the project for B. Tech does not contain any part of the work, which has been submitted for the award of any degree either in this University or in other University without proper citation.

Kanksha Masrani
(14IT058)

Prof. Jalpesh Vasa
Assistant Professor,
Department of Information Technology,
Faculty of Technology & Engineering,
Changa – 388421.

CHARUSAT

CERTIFICATE

This is to certify that the report entitled "**Prediction of Employee Attrition using Data Mining Techniques**" is a bonafied work carried out by **Ms. Kanksha Masrani (14IT058)** under the guidance and supervision of **Prof. Jalpesh Vasa** for the subject **Software Project Major (IT407)** of 8th Semester of Bachelor of Technology in **Information Technology** at Faculty of Technology & Engineering – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate herself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred to the examiner.

Under supervision of,

Prof. Jalpesh Vasa
Assistant Professor
Dept. of Information Technology
CSPIT, Changa, Gujarat.

Prof. Parth Shah
Head & Associate Professor
Department of Information Technology
CSPIT, Changa, Gujarat.

Chandubhai S Patel Institute of Technology

At: Changa, Ta. Petlad, Dist. Anand, PIN: 388 421. Gujarat

ABSTRACT

Employee turnover is a noteworthy matter in knowledge-based companies. On the off chance that employee leaves, they carry with them tacit information, often a source of competitive benefit to the other firms. Keeping in mind the end goal, to stay in the market and retain its employees, an organization requires to minimize employee attrition.

This project represents an illustration of employee churn forecast model utilizing established Machine Learning methods created. Model yields are then scrutinized to outline and experiment the best practices on employee withholding at different stages of the employee's association with an organization. This work has the potential for outlining better employee retention designs and enhancing employee contentment. The acknowledgment of the capability of Data Mining Techniques depends on the capacity to remove an incentive from huge information through information examination; this paper likewise incorporates and condenses the capacity to gain from information and give information-driven experiences, choice, and forecasts and thinks about significant machine learning systems that have been utilized to create predictive churn models.

ACKNOWLEDGEMENT

"It is not possible to prepare a project without the assistance & encouragement of other people. This one is certainly no exception."

On the very outset of this report, I would like to extend my sincere & heart felt obligation towards all the personages who have helped me in this endeavour. Without their active guidance, help, cooperation & encouragement I would not have made headway in the project.

I am extremely thankful and pay our gratitude to my faculty Prof. Jalpesh Vasa for his valuable guidance and support on completion of this project in it's presently. I am also extremely thankful and pay gratitude to our Head of Department Mr. Parth Shah for his valuable guidance and support on completion of this project. I extend my gratitude to CHANDUBHAI S. PATEL INSTITUTE OF TECHNOLOGY for giving us this opportunity.

At last but not least gratitude goes to all of my friends who directly or indirectly helped me to complete this project report. Any omission in this brief acknowledgement does not mean lack of gratitude.

Thanking You,

Kanksha Masrani (14IT058)

TABLE OF CONTENTS

List of Figures

List of Tables

ABBREVIATIONS

HR	**Human Resources**
HRIS	**HR Information Systems**
HRM	**Human Resource Management**
AUC	**Area Under the Curve**
ROC	**Receiver Operating Characteristic**
RF	**Random Forest**
SAS	**Statistical Analysis System**
BI	**Business Intelligence**

CHAPTER 1: INTRODUCTION

1.1 OVERVIEW

The beginning of globalization has provoked organizations and its pioneers to think and act internationally to be able to gain competitive advantage. Globally competitive organizations will rely on the uniqueness of their HR and the frameworks for managing HR adequately to end up productively. Indian associations are seeing a change in frameworks, administration societies and reasoning because of the global arrangement of Indian Associations.

HR speak to the aggregate aptitude, advancement, initiative, entrepreneurial and administrative abilities endowed in the employees of an organization. Every association involves individuals working in a particular department. Gaining their services, building up their aptitudes, motivating them to large amounts of execution, and guaranteeing that they keep up their responsibility regarding the association are basic to accomplishing hierarchical targets. Human Resource Management (HRM) is that piece of administration which is worried about individuals at work and with their connections inside an endeavour. It consists of practices that help the organization deal effectively with its people during the various phases of the employment cycle: pre-selection, selection, and post-selection. [1]

As Human Resources possesses a massive amount of employee data, demand for analyses is high. However, HR Information Systems (HRIS) are frequently underfunded contrasted with data frameworks of different areas of big business, which are straightforwardly associated with the principal business.[2] This prompts the way that HR information contains a ton of clamour and blunders. Therefore, building accurate analytical model is challenging for HR. One of the uses of Predictive Analytics for Human Resources (HR) is foreseeing employee turnover, attrition or retention. Employee turnover has various negative effects including loss of big business information, costs related to leaving and substitution.

Predictive analysis allows the analyst to operate on historical and current information as well as predicting the likely future environment. This predictive insight promotes much better decision making and improved results. Use of predictive analytics is wide, it enables companies to improve almost every aspect of their business.

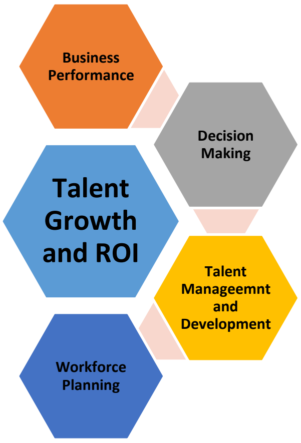


Fig 1.1 Usage of Predicitve Analytics

1.2 PROBLEM DEFINITION

The purpose of this study is to use Predictive Analytics for HR on example of employee turnover and to investigate variables that influence employee attrition within organization, using Machine Learning algorithms for the employee data. Aim is to try out different Machine Learning algorithms and evaluate their performance on company's data in order to select most accurate model. Accurate prediction of employee turnover will enable company to make strategic decisions regarding employee retention and take necessary actions.

1.3 MOTIVATION

With the development of modern information systems and databases that are capable of holding immense amounts of data, the need for analysing it has become progressively more relevant. This can be seen in trends of Google search, using keywords 'Big data', 'Data Science' and 'Machine Learning' (Figure 1.2). Also the fact that universities have started offering certification courses and master's degrees in Predictive Analytics and Big Data Analytics reflects the growth and popularity of this field. The raw data itself does not carry much value without any further processing and analysing. On Figure 1it is clearly visible that as Big Data became relevant problem for modern world, interest in Data Science and Machine Learning grew. There are different types of Data Analytics starting from Descriptive Analytics evolving to something more advanced, like Predictive Analytics

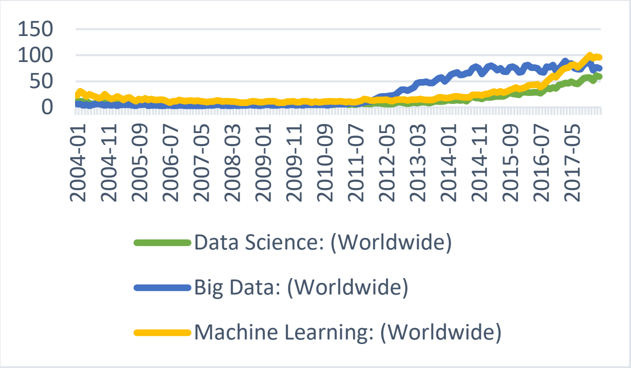


Fig 1.2 Word Trend [3]

1.4 SCOPE AND OBJECTIVE OF RESEARCH

The present study is aimed at identifying the critical factors affecting high employee attrition and suggesting remedial measures to address the high attrition problem. The different sectors (area of work) included in this study are financial accounting, customer services, procurement, human resource, application process and others. The study gives a warning signal to the sectors, to immediately adopt innovative strategies to tackle the continuing high attrition problem.

This study will be helpful to the management to focus on the critical factors identified in the study in addressing the employee attrition problem. Also the study will enable the readers, researchers and practitioners (HR Managers) to have a professional approach in addressing the critical issue of employee attrition.

Objective of the study:

- To study the variation in factors causing high employee attrition in various departments.
- To study the relationship between the parameters included and its relation with employee turnover.
- To identify the best Machine Learning algorithm suitable for the dataset taken and find out the best algorithm on the basis on accuracy, time and F1 Score.
- To suggest innovative measures for reducing employee attrition.

CHAPTER 2: LITERATURE REVIEW

2.1 REVIEW OF LITERATURE AND FINDINGS

1) "Towards Applying Data Mining Techniques for Talent Managements", 2009

Table 2.1 Literature review 1

PROBLEM INCLUDED	Data Mining techniques for performance prediction of employees
AUTHOR	Jantan, Hamdan and Othman[4]
DATA MINING TECHNIQUE STUDIED	C4.5 decision tree, Random Forest, Multilayer Perceptron(MLP) and Radial Basic Function Network
RECOMMENDED	C4.5 decision tree

2) "Establishing a link between employee turnover and withdrawal behaviours: Application of data mining techniques",2008

Table 2.2 Literature review 2

PROBLEM INCLUDED	Relationship of withdrawal behaviors like lateness and absenteeism, job content, tenure and demographics on employee turnover
AUTHOR	Nagadevara, Srinivasan and Valk[5]
DATA MINING TECHNIQUE STUDIED	Artificial neural networks, logistic regression, classification and regression trees (CART), classification trees (C5.0), and discriminant analysis)
RECOMMENDED	Classification and regression trees (CART)

3) “A comparative test of two employee turnover prediction models”, 2007

Table 2.3 Literature review 3

PROBLEM INCLUDED	Feasibility of applying the Logit and Probit models to employee voluntary Turnover predictions.
AUTHOR	Hong, Wei and Chen[6]
DATA MINING TECHNIQUE STUDIED	Logistic regression model (logit), probability regression model (probit)
RECOMMENDED	Logistic regression model (logit)

4) “Predictive Models of Employee Voluntary Turnover in a North American Professional Sales Force using Data-Mining Analysis”, 2007

Table 2.4 Literature review 4

TITLE & PROBLEM INCLUDED	To explore various personal, as well as work variables impacting employee voluntary turnover
AUTHOR	Marjorie Laura Kane-Sellers[7]
DATA MINING TECHNIQUE STUDIED	Binomial logit regression
RECOMMENDED	Binomial logit regression

5) “Analyzing employee attrition using decision tree algorithms”, 2013

Table 2.5 Literature review 5

PROBLEM INCLUDED	Analyzing employee attrition using multiple decision tree algorithms
AUTHOR	Alao and Adeyemo[8]
DATA MINING TECHNIQUE STUDIED	C4.5, C5, REPTree, CART
RECOMMENDED	C5 decision tree

6) “Employee churn prediction”, 2011

Table 2.6 Literature review

PROBLEM INCLUDED	To compare data mining techniques for predicting employee churn
AUTHOR	Saradhi and Palshikar[9]
DATA MINING TECHNIQUE STUDIED	Naïve Bayes, Support Vector Machines, Logistic Regression, Decision Trees and Random Forests
RECOMMENDED	Support Vector Machines

CHAPTER 3: PROPOSED DESIGN

3.1 INTRODUCTION TO THE PROPOSED SYSTEM

3.1.1 Modelling Process

Building a Predictive Analytics solution is a continuous and an iterative process which requires an integrated enterprise wide approach. Since business objectives and related datasets vary company to company Predictive Analysis process might be somewhat different, but it can be structured as six steps approach shown on Figure 3.1.

1. Identify the business objective - Defining outcomes and business objectives is very first step for building a Predictive Analytics project.

2. Prepare the data - Data collection and data analysis are required for preparing data, to be later used in predictive modelling. Data should be cleaned and transformed, so that useful information can be concluded. Statistical analysis can be applied to validate assumptions, by testing using standard statistical models. Data Mining can be used for data preparation from multiple sources.

3. Develop the Predictive model - Predictive modelling provides ability to automatically create accurate predictive models about future. Predictive modelling is process of creating, testing and validating a model to best predict the probability of outcome. Modelling methods are for example: Machine Learning, AI, statistics.

4. Test the model

5. Deploy the model - Predictive model deployment provides the option to deploy the analytical results into the everyday decision making process to get results, reports and output by automating the decisions based on the modelling.

6. Monitor for effectiveness - Models are managed and monitored to review the model performance to ensure that it is providing results expected.

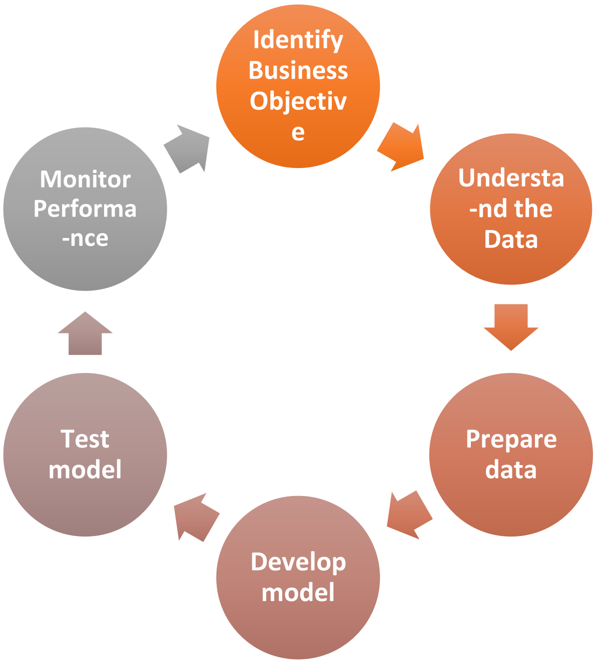


Fig 3.1 Modelling Process

3.2 ALGORITHM OF PROPOSED SYSTEM

3.2.1 Developing predictive model

Most algorithms used for predictive analyses have been out there for decades, but only recently data scientists started to mine data effectively. Since, just recently data gathering has become cheaper and faster. Algorithm that should be used for data modelling should be determined and developed.

3.2.2 Learning methods

There are two main Machine Learning types: **supervised and unsupervised** [10].

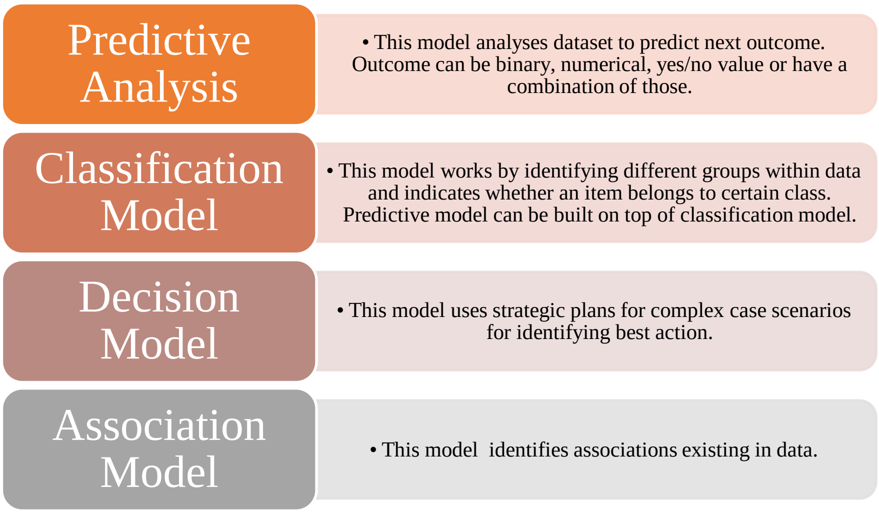
In case of **supervised** learning, machine is told what the correct answers are, input training data is labelled and results are known. Machine is corrected when wrong prediction is made, so it learns on previous experience. Supervised learning method is suitable for both classification and regression problems.

When data is not labelled and results are unknown **unsupervised** learning method is used. Such algorithms are used for clustering and association problems.
There can be mixture of both labelled and unknown data as well. Such learning style is called semi-supervised learning. For such learning methods, algorithm should organize data and then predict.

In addition to those three there is also reinforcement learning, which uses reward function for training. Essentially in reinforcement learning an agent will be rewarded as it transitions to certain states or executes actions. The algorithm tries to optimize its actions for maximizing the reward. For example, a Machine Learning algorithm can be taught to play a computer game, if the agent has a control over the inputs and is rewarded based on the game score, resulted from the correct sequence of inputs. [11]

3.2.3 Different Model categories

Most Predictive Analytics tools include algorithms for modelling. Building a model with best accuracy possible will take more time. Experiment with different approaches should be held and outcomes should be re-evaluated. Model requires continuous updates to keep relevance, building a predictive model should be an on-going process. Predictive Analysis models can be categorised various ways. For example, they can be categorised by different approaches like:



Predictive Analysis

- This model analyses dataset to predict next outcome. Outcome can be binary, numerical, yes/no value or have a combination of those.

Classification Model

- This model works by identifying different groups within data and indicates whether an item belongs to certain class. Predictive model can be built on top of classification model.

Decision Model

- This model uses strategic plans for complex case scenarios for identifying best action.

Association Model

- This model identifies associations existing in data.

Fig 3.2 Model Types

3.3 WORKING OF THE PROPOSED SYSTEM

A set of algorithms and tools gathered under an umbrella term ‘Machine Learning’ will be concentrated on, as they have shown the most promise predicting on large datasets. Machine Learning Algorithms will be applied on the processed Dataset and will be ranked on the basis of various factors such as accuracy and F1 score. The optimal algorithm will be chosen among those and will be considered for further prediction.

The flow of the project will be as shown below:

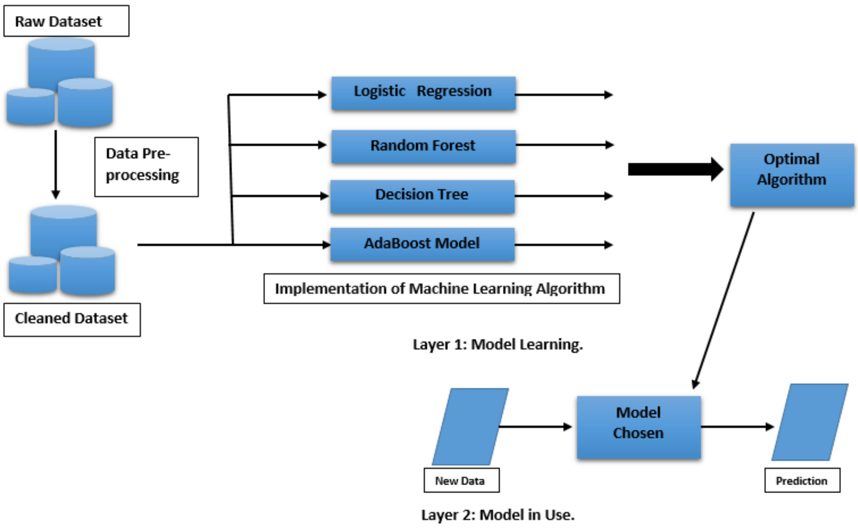


Fig 3.3 Flow of Project

3.3.1 Methods for Predictive Analysis

Machine learning uses computer program that learns by analysing data. The more data there is, the better the program can learn. We are teaching the program rules and program itself is getting better at it by learning and practicing. Machine learning algorithms are more suitable for complex data and work better with huge sets of training examples []. There are many different types of Machine learning algorithms. On Figure 3.4. Machine Learning algorithms are categorised according to methodologies they use for prediction.

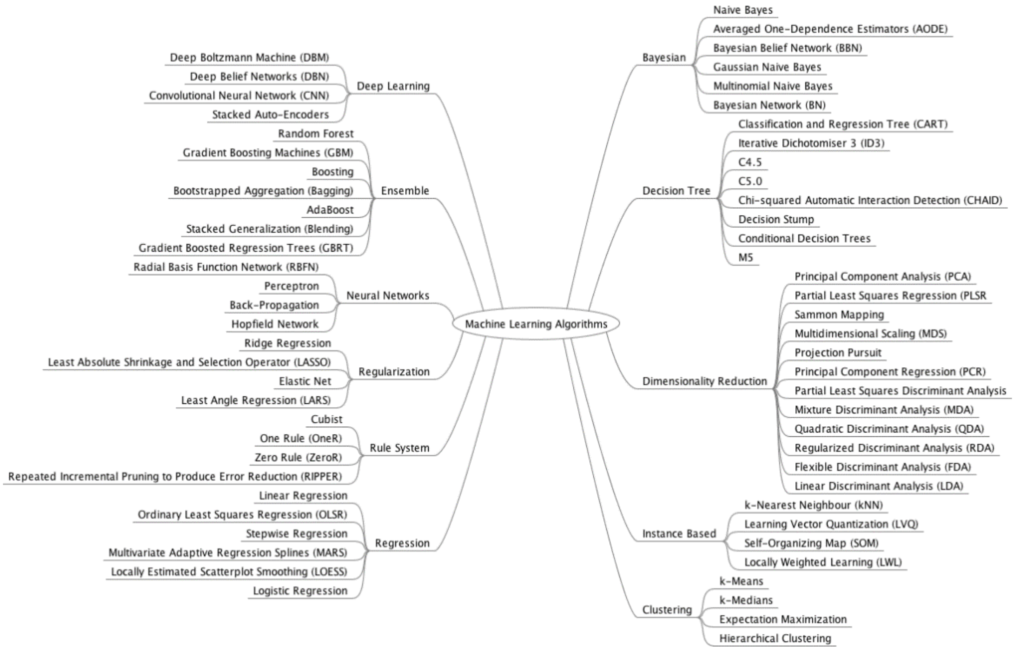

Machine Learning Algorithms
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering

Fig 3.4 Machine Learning Algorithms [12]

The following Machine Learning Algorithms have been implemented:

Logistic Regression:

Logistic regression predicts the likelihood of a result that can just have two esteems (i.e. a dichotomy). The forecast depends on the utilization of one or a few indicators (numerical and categorical). A linear regression isn't suitable for anticipating the estimation of a binary variable for the accompanying reason:

A linear regression will anticipate values outside the worthy range (e.g. Anticipating probabilities outside the range 0 to 1). Since the dichotomous examinations can just have one of two conceivable esteems for each test, the residuals won't be typically distributed about the anticipated line.

Then again, a logistic regression produces a logistic curve, which is restricted to values in the vicinity of 0 and 1. Logistic regression is similar to a linear regression, however the curve is developed utilizing the common logarithm of the "chances" of the objective variable, as opposed to the likelihood. Besides, the indicators don't need to be normally distributed or have equal variance in each group.

Decision tree:

Decision trees are graphical portrayals of elective decisions that can be made by a business, which empower the chief to distinguish the most reasonable alternative in a specific situation. Decision trees will be trees that group examples by arranging them based on feature values. Every node in a decision tree speaks to an element in an instance to be classified, and each branch speaks to a value that the node can assume. Cases are characterized beginning at the root hub and arranged in light of their feature value. [13]

Greedy algorithm is a fundamental calculation for decision tree induction that develops the tree in a top-down recursive divide-and-conquer way. This algorithm is generally utilized on the grounds that they are proficient and simple to execute, yet they usually lead to sub-optimal models. Another alternative approach can be a bottom-up approach.

Decision trees utilized as a part of data mining are of two principle types:

- Classification tree analysis is when the predicted outcome is the class to which the data belongs.
- Regression tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient’s length of stay in a hospital).

Adaptive Boosting:

Boosting refers to the general problem of producing a very accurate prediction rule by combining rough and moderately inaccurate rules-of-thumbs. [14] This includes fitting a sequence of weak learners on changed data. The predictions from every one of them are then combined through greater part vote (or sum) to deliver the last prediction. The data alteration at each progression comprises of doling out higher weights to the training examples that were misclassified in the previous iteration.

As iteration continues, cases that are hard to foresee get regularly expanding influence. This forces the weak learner to focus on the cases that are missed by its antecedent. AdaBoost is a boosted tree algorithm. It follows the principle of gradient boosting [15] In comparison to gradient boosting, it makes use of a more standardised-model formalization to control over-fitting, giving it a better execution.

Random Forest:

Random Forest calculation is a popular tree-based ensemble learning method. The sort of 'ensembling' utilized here is bagging. [16] In bagging, progressive trees don't rely upon prior trees — each is independently constructed utilizing an alternate bootstrap sample of the data set. At last, a basic greater part vote is taken for forecast. Random forest are not quite the same as standard trees in that for the latter each node is split utilizing the best split among all factors. In a random forest, each node is split utilizing the best among a subset of indicators randomly picked at that node [17] this extra layer of randomness makes it robust against over-fitting. [18]

3.3.2 Tools for Predictive Analysis

There are various advanced tools and software existing, that are used for Predictive Analytics. These tools vary vendor to vendor, with functionality, different type of usage, customization, etc. Mostly Predictive Analytics tools are used in marketing, for customer classification, customer churn and so on. Many Banks and financial companies also use Predictive Analysis for fraud detection and risk management. Various industries use Predictive Analytics for sales forecasting, sentiment analysis, predicting employee performance, identifying patters in behaviour and so on. Within this various and numerous tools available in marketplace, some are free and open source and some are proprietary, there is large number of APIs as well. Well know open source tools are: R, WEKA, Rapid Miner, NumPy and many others. SAS, SAP, IBM Predictive Analytics, STATISTICA, Matlab would be examples of proprietary software which can be used for predictive modelling.

Technology that is used for Predictive Analysis should include Data Mining capabilities, Statistical methods, Machine Learning algorithms and software tools for building the model. Choosing optimal tool for usage can be a hard task. Companies choose right tool for predictive analyses based on the price of the tool, complexity of data or business goal, data source, data growth speed, people skills who should use product and etc. [19]

In this case Python has been used for all the Analysis and for Data Visualization Public Tableau has been used.

Python

Python is a general-purpose programming language that is also widely used for statistical computing and Machine Learning. Even though it was not developed with mathematics or data analysis directly in mind, the community has created numerous libraries like NumPy [20], Pandas [21], Scikit-learn [22], Tensorflow [23], which improve on mathematical and statistical operations, data structures add numerous Data Analytics capabilities and Machine earning algorithms.

The main advantage of using a general-purpose programming language for Data Analytics is the flexibility of it and possibility of easily extending on it. This also means that deploying the code in production is feasible, especially if rest of the project also makes use of python. As in the Machine Learning community python is becoming the de-facto programming interface and it is becoming more and more searched and requested.

Since Python is an open tool and is free for users, other parties can design their own packages and extend pythons functionality. Many users of python can contribute also by reporting issues or making small improvements in to the code [24].

Public Tableau

General visualization tools typically require manual specification of views: analysts must select data variables and then choose which transformations and visual encodings to apply. These decisions often involve both domain and visualization design expertise, and may impose a tedious specification process that impedes exploration.

Since Tableau provides various visualizations and customizations, the level of analysis can be increased with small multiples, view filtering, mark cards, and Tableau charts. Conclusions Tableau is a software that can help users explore and understand their data by creating interactive visualizations. The software has the advantages that it can be used in conjunction with almost any database, and it is easy to use by dragging and dropping to create an interactive visualization expressing the desired format.

CHAPTER 4: IMPLEMENTATION DETAILS

4.1 BUILDING THE PREDICTIVE MODELS

Since showing any company data is a security issue, to show data preparation dataset from Kaggle – Human Resource Analytics has been used for analysis[]. Example dataset contains information about employees of a company including attrition attribute. Outcome of the analysis should be the probability of attrition. In example dataset there are 8 variables, including the outcome variable (attrition). The dataset has 15000 data observations in data file.

To predict employee turnover, we should predict who will leave. From the total, 24% left the company and 76% stayed. That is binomial classification problem, where set of employees should be divided into two groups based on some characteristics, ones with higher risk of attrition and ones with less.

Commonly used Machine Learning algorithms for binomial classification problem are: Decision Trees, Random Forest, Boosting, Bayesian networks, Neural Networks, Support Vector Machine and Logistic Regression, out of which below will be used Random Forest, Logistic Regression, Decision trees and Adaptive Boosting.

4.1.1 Data Pre-processing

Raw data is highly susceptible to noise, missing values, and inconsistency. The quality of data affects the data mining results. In order to help improve the quality of the data and, consequently, of the mining results raw data is pre-processed so as to improve the efficiency and ease of the mining process. Data pre-processing is one of the most critical steps in a data mining process which deals with the preparation and transformation of the initial dataset.

Data that is to be analyzed by data mining techniques can be incomplete, noisy, and inconsistent. Incomplete, noisy, and inconsistent data are commonplace properties of large, real world database and data warehouse. The following steps were performed:

- Missing values were replaced by taking the attribute value.
- Conversion of data into categorical data, such as the “department” and “Salary” features.
- Renaming features for better readability.

The feature “turnover” is considered our dependent variable, whereas the rest of the features were our independent variable.

The following independent variables were used in the model:

- **Satisfaction**: An employee’s level of satisfaction in percentage
- **Evaluation:** An employee’s evaluation score in percentage
- **Project Count:** The amount of projects the employee has done
- **Average Monthly Hours:** The total monthly hours an employee worked
- **Years At Company:** The number of years an employee was at the company
- **Work Accident:** Whether an employee had an accident or not. Where 0 (zero) means no and 1 (one) means yes
- **Promotion:** Whether an employee had a promotion within the last five years. Where 0 (zero) means no and 1 (one) means yes.
- **Department:** The type of department an employee worked under. Which includes sales, accounting, hr, technical, support, management, IT, product management, and marketing.
- **Salary:** The type of salary an employee got, which ranges from low, medium, or high.

4.1.2 Exploratory Data Analysis

a) Statistical Overview

Here are some important numbers to keep in mind of the dataset:

- There is 15,000 employees and 9 independent variables
- Turnover rate: 24%
- Mean satisfaction: 0.61

b) Correlation Matrix & Heat map

From the heat map, there is a positive (+) correlation between the variables: project Count, AverageMonthlyHours, and evaluation. Which means that the employees who spent worked more hours and did more projects had higher evaluations. For the negative (-) relationships, the most important feature that correlated with our target variable (turnover) is satisfaction. This should support our initial intuition that employees who tend to quit would normally have lower satisfaction level.

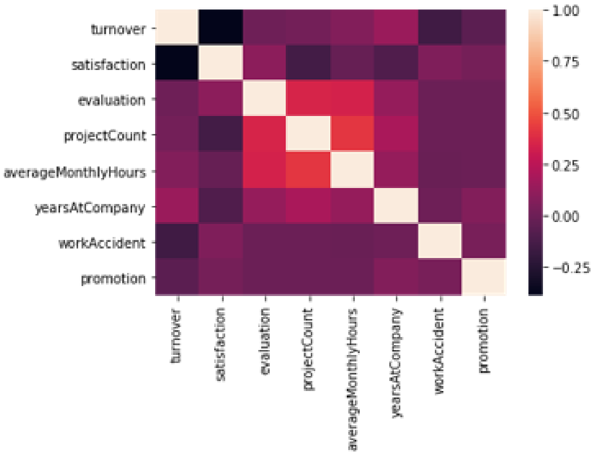


Fig 4.1 Heat Map (Correlation)

c) Histogram (Satisfaction/Evaluation/Average Monthly Hours)

Next comes examining the distribution on some of the employee's features:

-Satisfaction: There are three distributions for employee satisfaction in the dataset. One group falls within satisfaction level of (0–0.3), another within satisfaction level of (0.3-0.5), and one from (0.5-1).

-Evaluation: There are three distributions for employee evaluation. One group falls within the lower spectrum of (0-0.55), the middle spectrum of (0.55-.7), and the higher spectrum of (.7-1).

-**Average Monthly Hours:** There are three distributions for employee average monthly hours. Those that work 100-150 hours, another group that works 150-250 hours, and a third group who works 250-300 hours.

-The **evaluation** and **average monthly hour's** features both share a similar distribution, which indicates high collinearity of the features.

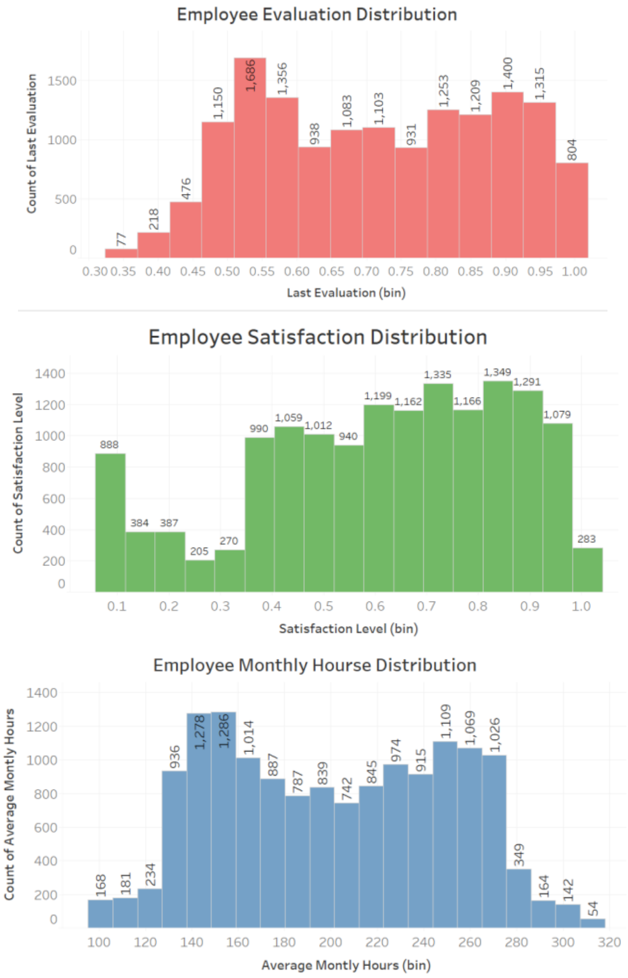


Fig 4.2 Distribution of Employees

d) Salary V/S Turnover

Barely any employee with high salary left. Majority of employees left were a part of the low and medium salary range.

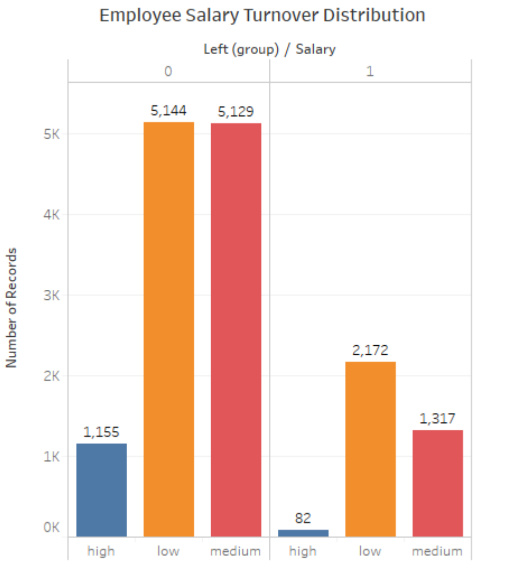


Fig 4.3 Employee Salary – Turnover Distribution

e) Department V/S Turnover

When analysis was done department wise the sales, technical and support department had the highest employee turnover whereas the management department has the least percentage of turnover.

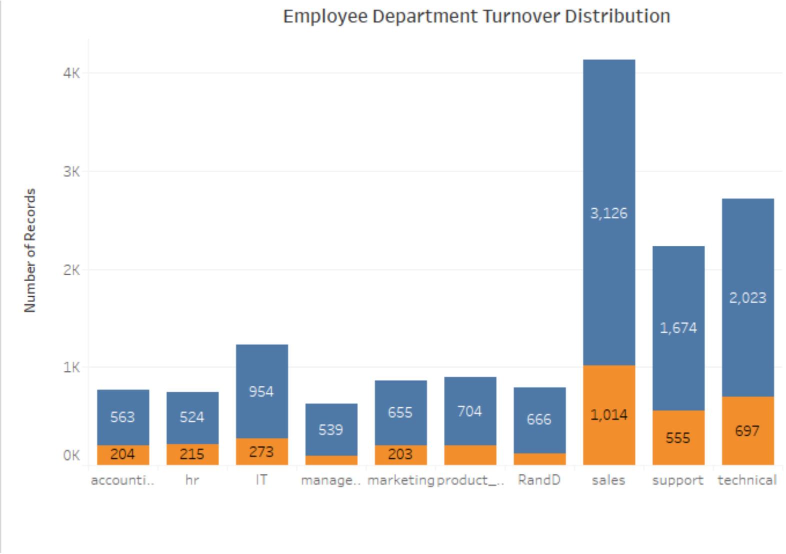


Fig 4.4 Employee Department – Turnover Distribution

f) Project Count V/S Turnover

More than half of employees with 2, 6, and 7 projects left the company Majority of the employees who did not leave had 3,4, and 5 projects and all employees with 7 projects left the company Hence, there is an increase in turnover as project count increases

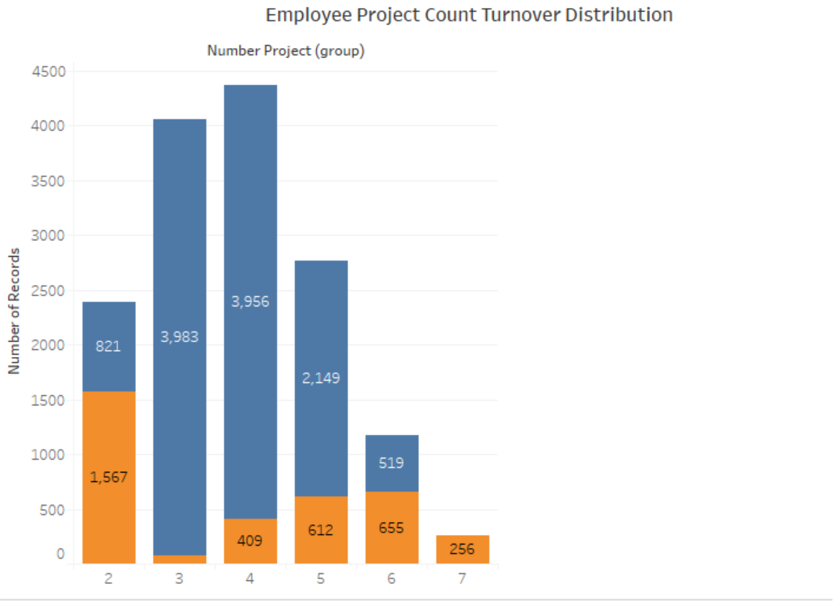


Fig 4.5 Employee Project Count – Turnover Distribution

g) Years at Company V/S Turnover

After analysis it was seen that more than half of the employees with 4 and 5 years left the company.

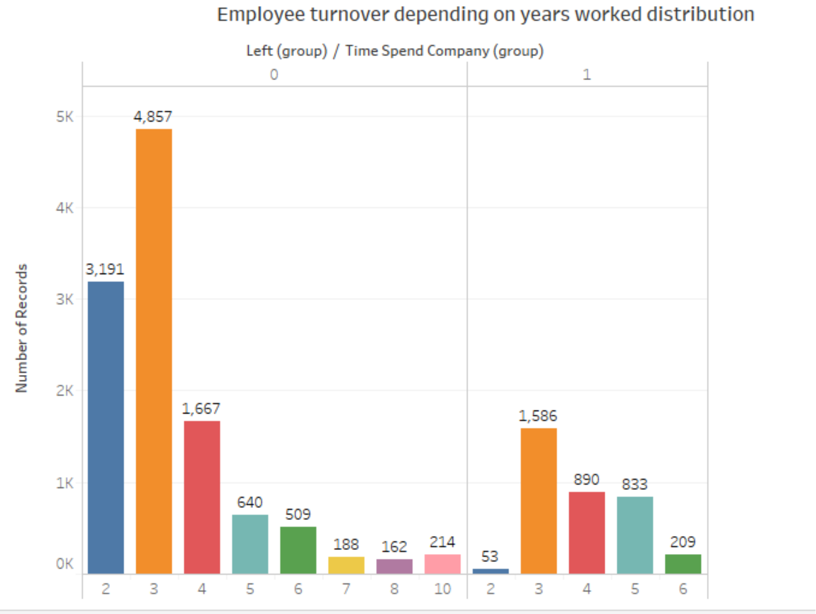


Fig 4.6 Employee Years at Company – Turnover Distribution

h) Evaluation V/S Turnover

There is a bimodal distribution for employees that left the company:
-Employees with low evaluation levels (0.2-0.6) and high evaluation levels (0.8-1) were the bulk of employee turnover
-Employees with evaluation levels (0.6-0.8) had the smallest turnover rate

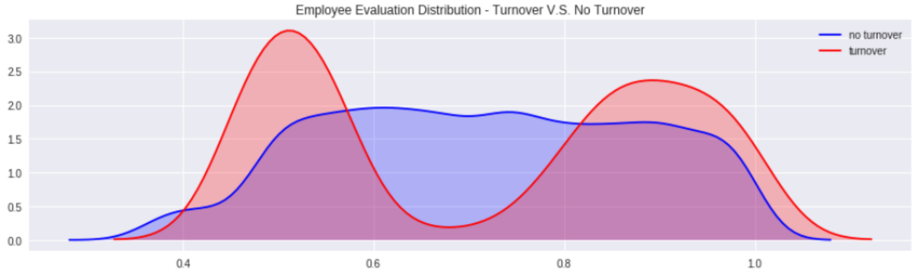


Fig 4.7 Employee Evaluation – Turnover Distribution

i) Satisfaction V/S Turnover

There is a tri-modal distribution for employees that left the company
-Employees left with really low satisfaction levels of (0-0.2)
-Employees left with low satisfaction levels of (0.3-0.5)
-Employees left with high satisfaction levels of (0.7-1)

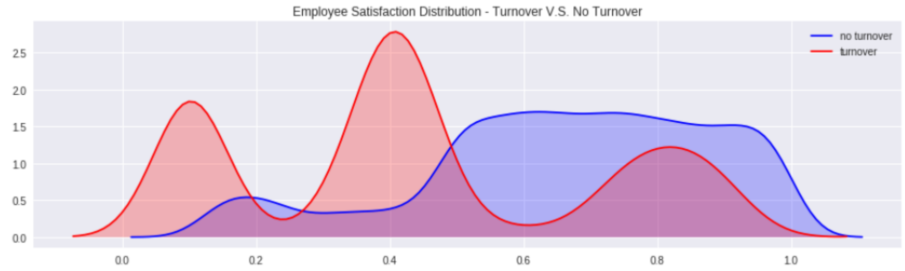


Fig 4.8 Employee Satisfaction– Turnover Distribution

j) Average Monthly Hours V/S Turnover

There is another bimodal distribution for employees that left the company
-Employees who had less hours of work (~150 hours or less) left the company more
-Employees who had more hours of work (~250 hours or more) left the company more
-Employees who left generally were underworked or overworked

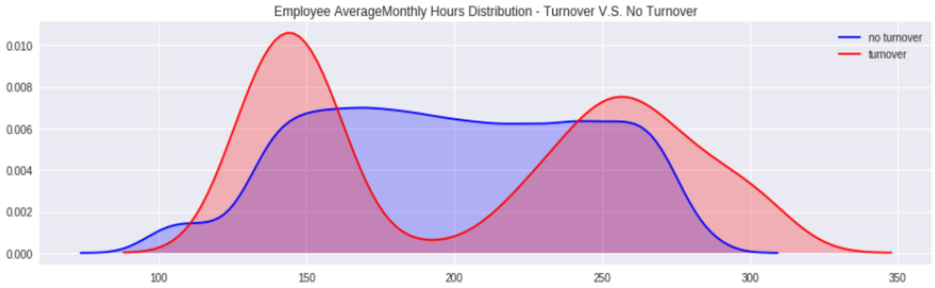


Fig 4.9 Employee AvgMonthlyHrs – Turnover Distribution

k) Evaluation V/S Satisfaction

After plotting graphs between the attributes depicting years at company to the turnover most employees left after working for 4-5 years. An Interesting insight was drawn from plotting of satisfaction and evaluation attributes by creating clusters. It can be depicted in the table below:

Table 4.1 Evaluation v/s Satisfaction Cluster

Satisfaction	Evaluation	Cluster
< 0.2	>0.75	Overworked
0.35~0.45	~0.58	Under-performed
0.7~0.1	>0.8	Ideal

Table 4.2 Conclusion Derived

Cluster	Conclusion
Overworked	Good employees left who felt horrible at work.
Under-performed	Badly evaluated and felt horrible at work.
Ideal	Loved their work and had a great performance. Left due to better job opportunities.

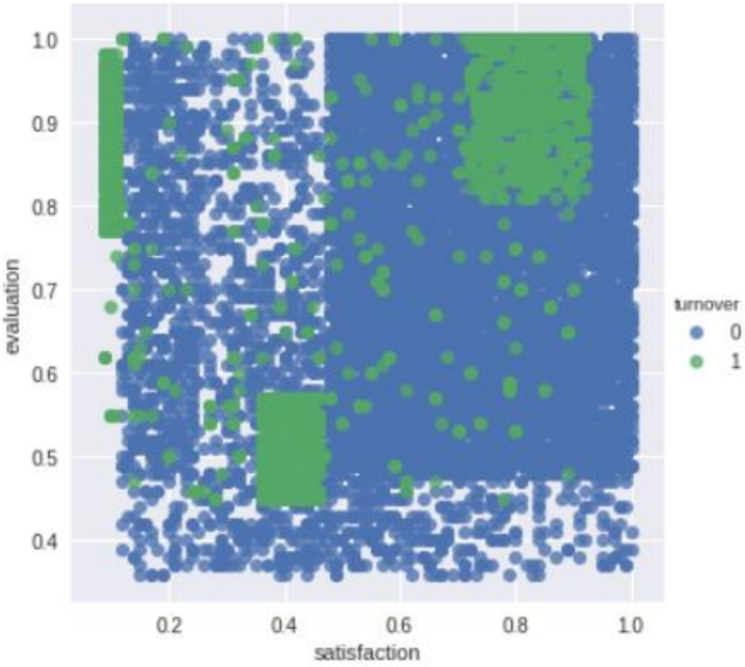


Fig 4.10 Employee Evaluation – Satisfaction Clustering

4.1.3 Modeling on the dataset

After pre-processing the data, four different Machine Learning models have been applied. Logistic Regression, Decision Tree, Random Forest & Adaptive Boosting were selected algorithms, based on different approaches they use for prediction.

Each model first was configured and most suited parameters were chosen. For evaluation sklearn.cross_validation.train_test_split was used. Dataset was randomly split so that 70% was used for training and 30% for validation. Train-test split was repeated several times and mean model accuracy was calculated.

By using a decision tree classifier, it could rank the features used for the prediction. The top three features were employee satisfaction, yearsAtCompany, and evaluation. This is helpful in creating our model for logistic regression because it'll be more interpretable to understand what goes into our model when we utilize less features.

For evaluating model performances sklearn.metrics.roc_auc_score function was used, which computes Area under the Curve (AUC) from prediction scores, the receiver operating characteristic (ROC) curve. Output of model was prediction of attrition for each employee, along with prediction probability. For calculating prediction probabilities predict_proba function was used.

For evaluating model performances it is more convenient to use one metric.
Confusion matrix of the dataset for prediction model will be:

		Predicted class	
		Employee	**Leaver**
Actual Class	**Employee**	**e**	**`e**
	Leaver	**`l**	**L**

Table 4.3 Confusion matrix

Where e is number of correctly classified data object (called true positive), in this study case correctly predicted number of employees. $\bar{e}$ is number of misclassified employees (false negatives). Analogically, $\bar{l}$ is number of misclassified leavers (false positive) and l is number of correctly predicted leavers (true negative).
From confusion matrix two metrics can be calculated:

$True\ positive\ rate = e \,/\, e + \bar{e}$

$False\ positive\ rate = \bar{l} \,/\, \bar{l}+l$

ROC curve is plotted using true positive rate and false positive rate coordinates. AUC calculates the area under the ROC curve.

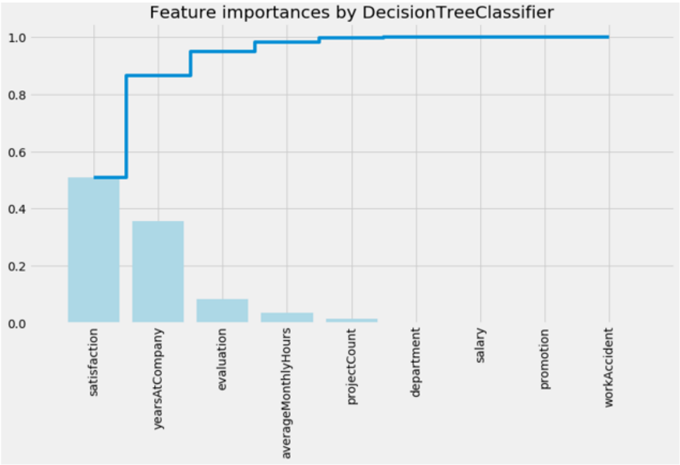


Fig 4.11 Decision Tree Classifier

4.1.4 PSEUDO CODE

```
#Code for Decision Tree Model
import time
start_time=time.clock()
from sklearn.metrics import roc_auc_score
from sklearn.metrics import classification_report
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.ensemble import BaggingClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.ensemble import VotingClassifier
import time

# Decision Tree Model
dtree = tree.DecisionTreeClassifier(
    #max_depth=3,
    class_weight="balanced",
    min_weight_fraction_leaf=0.01
    )
dtree = dtree.fit(X_train,y_train)
print ("\n\n ---Decision Tree Model---")
dt_roc_auc = roc_auc_score(y_test, dtree.predict(X_test))
end_time = time.clock()
time_dtree= end_time-start_time
print(time_dtree)
print ("Decision Tree AUC = %2.2f" % dt_roc_auc)
```

```
print(classification_report(y_test, dtree.predict(X_test)))
print(confusion_matrix(y_test,dtree.predict(X_test)))
```

```
# Code for Random Forest
import time
start_time= time.clock()
from sklearn.metrics import roc_auc_score
from sklearn.metrics import classification_report
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.ensemble import BaggingClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.ensemble import VotingClassifier
import time

# Random Forest Model
rf = RandomForestClassifier(
    n_estimators=1000,
    max_depth=None,
    min_samples_split=10,
    class_weight="balanced"
    #min_weight_fraction_leaf=0.02
    )
rf.fit(X_train, y_train)
print ("\n\n ---Random Forest Model---")
rf_roc_auc = roc_auc_score(y_test, rf.predict(X_test))
end_time = time.clock()
time_RF = end_time-start_time
print(time_RF)
print ("Random Forest AUC = %2.2f" % rf_roc_auc)
print(classification_report(y_test, rf.predict(X_test)))
print(confusion_matrix(y_test,rf.predict(X_test)))
```

```
# Code for ROC Graph
from sklearn.metrics import roc_curve
fpr, tpr, thresholds = roc_curve(y_test, logis.predict_proba(X_test)[:,1])
rf_fpr, rf_tpr, rf_thresholds = roc_curve(y_test, rf.predict_proba(X_test)[:,1])
```

```
dt_fpr, dt_tpr, dt_thresholds = roc_curve(y_test, dtree.predict_proba(X_test)[:,1])
ada_fpr, ada_tpr, ada_thresholds = roc_curve(y_test,
ada.predict_proba(X_test)[:,1])

plt.figure()

# Plot Logistic Regression ROC
plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit_roc_auc)

# Plot Random Forest ROC
plt.plot(rf_fpr, rf_tpr, label='Random Forest (area = %0.2f)' % rf_roc_auc)

# Plot Decision Tree ROC
plt.plot(dt_fpr, dt_tpr, label='Decision Tree (area = %0.2f)' % dt_roc_auc)

# Plot AdaBoost ROC
plt.plot(ada_fpr, ada_tpr, label='AdaBoost (area = %0.2f)' % ada_roc_auc)

# Plot Base Rate ROC
plt.plot([0,1], [0,1],label='Base Rate' 'k--')

plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Graph')
plt.legend(loc="lower right")
plt.show()
```

```
# Decision Tree Classifier
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
plt.style.use('fivethirtyeight')
plt.rcParams['figure.figsize'] = (12,6)

# Renaming certain columns for better readability
df = df.rename(columns={'satisfaction_level': 'satisfaction',
                        'last_evaluation': 'evaluation',
                        'number_project': 'projectCount',
                        'average_montly_hours': 'averageMonthlyHours',
                        'time_spend_company': 'yearsAtCompany',
                        'Work_accident': 'workAccident',
                        'promotion_last_5years': 'promotion'
```

```
                'sales' : 'department',
                'left' : 'turnover'
                })

# Convert these variables into categorical variables
df["department"] = df["department"].astype('category').cat.codes
df["salary"] = df["salary"].astype('category').cat.codes

# Create train and test splits
target_name = 'turnover'
X = df.drop('turnover', axis=1)


y=df[target_name]

X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.15,
random_state=123, stratify=y)

dtree = tree.DecisionTreeClassifier(
   #max_depth=3,
   class_weight="balanced",
   min_weight_fraction_leaf=0.01
   )
dtree = dtree.fit(X_train,y_train)

## plot the importances ##
importances = dtree.feature_importances_
feat_names = df.drop(['turnover'],axis=1).columns


indices = np.argsort(importances)[::-1]
plt.figure(figsize=(12,6))
plt.title("Feature importances by DecisionTreeClassifier")
plt.bar(range(len(indices)), importances[indices], color='lightblue',
align="center")
plt.step(range(len(indices)), np.cumsum(importances[indices]), where='mid',
label='Cumulative')
plt.xticks(range(len(indices)), feat_names[indices], rotation='vertical',fontsize=14)
plt.xlim([-1, len(indices)])
plt.show()
```

CHAPTER 5: EXPERIMENTAL RESULTS

Application of fundamental statistical strategies is utilized to study the employee. Furthermore, the model runtime is additionally used to analyse the execution of the classifiers. This measure is vital to be considered, as it fabricates a case from an expert's point of view on figuring out the algorithm which is worthy to implement for real-life business issues, answering for adaptability and execution. ROC curve was plotted using true positive rate and false positive rate coordinates. AUC calculates the area under the ROC curve.

5.1 ROC Curve

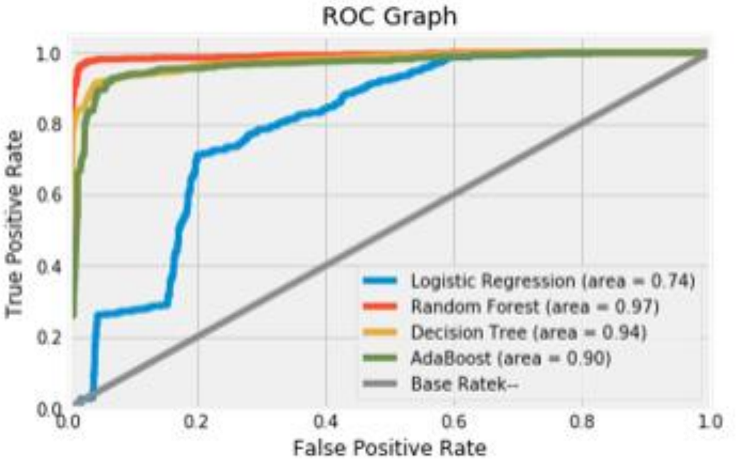


Fig 5.1 ROC Graph

5.2 Comparison based on various Features

A comparison between all the four Machine Learning Algorithms was generated with respect to various features mentioned and contingency matrix.

Table 5.1 Algorithm Comparison Table

Features	Logistic Regression	Decision Tree C4.0	Adaptive Boosting	Random forest
Simplicity	Very Simple	Moderate	Moderate	Difficult
Performance	Average	Good	Good	Excellent
Accuracy	0.7489	0.9476	0.9356	0.9773
Time required for training classifier	Less (0.021 sec)	Moderate (0.167 sec)	High (2.515 sec)	Recurrent Learning with every novel dataset
AUC/ROC	0.74	0.94	0.90	0.97

5.3 Contingency Table $\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$

A Binary classification model classifies each instance into one of the two classes; say a true and a false class. This gives rise to four possible classifications for each instance: a true positive, a true negative, a false positive, or a false negative. This situation can be depicted as a confusion matrix (also known as contingency table) given in Table 5.2. The confusion matrix juxtaposes the observed classifications for a phenomena with the predicted classification of a model.

Table 5.2 Contingency Table

	Logistic Regression	**Decision Tree**	**Adaptive Boosting**	**Random Forest**
True Positive (TP)	1296	1642	1664	1686
True Negative (TN)	389	490	441	513
False Positive (FP)	418	72	50	28
False Negative (FN)	147	46	95	23
Recall	0.75	0.95	0.94	0.98
Precision	0.80	0.95	0.93	0.98
F1 Score	0.76	0.95	0.93	0.98
Accuracy	0.7489	0.9476	0.9356	0.9773

Precision/Positive Predictive value = $TP/TP + FP$
Negative Predictive value = $TN/FN +$TN
Sensitivity = $TP/TP + FN$
Specificity= $TN/FP + TN$
Accuracy = $TP + TN/TP + FP + FN + TN$

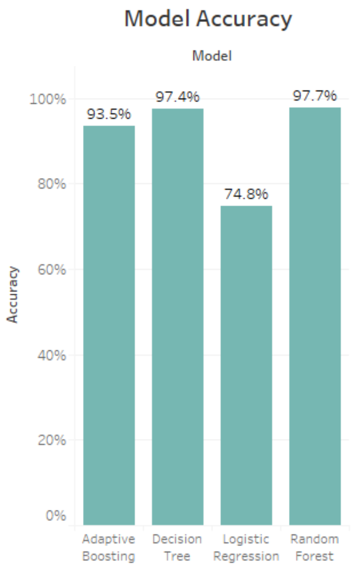


Fig 5.2 Model Accuracy Distributio

5.4 Model Verdict.

The machine learning models performed at varying performance levels for the different input datasets. By considering the average F5 score, model rankings are obtained- i.e. a higher F1 score is indicative of a better performing model.

Model Name	Avg. F1 score	Ranking
Base Model	0.66	5
Logistic Model	0.76	4
Adaptive Boosting Model	0.93	3
Decision Tree Model	0.95	2
Random Forest Model	0.98	1

Table 5.3

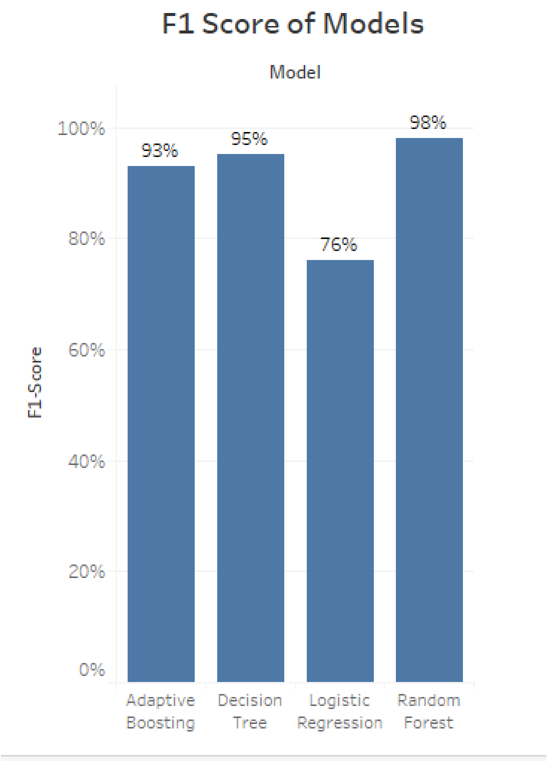


Fig 5.3 F1 Score Distribution

CHAPTER 6: CONCLUSION

The aim of the project was to apply Predictive Analysis methodology on HR data, in order to analyse the use of predictive analysis for HR. Chosen metric for prediction was employee turnover, because of its high importance for organization. Thus goal of the project was predicting employee turnover.

In this work, data preparation has been done on the dataset available on Kaggle. Example dataset was used for demonstrating methods used for cleaning and preparing data. Then missing values has been imputed, outliers have been removed and skew has been reduced. Afterwards parameters for selected Machine Learning algorithms have been tuned using different methodologies. Later, on pre-processed employee data different machine learning algorithms have been applied and algorithm with best prediction accuracy has been selected. In addition, for understanding how decision was made, the model was interpreted by plotting decision graph and identifying most important features.

Outcome of the project is employee turnover prediction. The aim is reached as application of model was successful. To conclude, Random Forest classifier performed better than other tested models based on accuracy, timing and F1 score. Furthermore, feature importance has been evaluated to better understand variables that influenced the decision most.

CHAPTER 7: FUTURE ENHANCEMENTS

For future examinations, I propose the capture of information around interventions done by the organization for in danger employees and their result. This will change the model into a normative one, addressing not only "Who is at risk?" but also "What would we be able to do?"

It is also recommended to study the application of deep learning models for anticipating turnover. A very much outlined system with adequate hidden layers may enhance the precision, be that as it may, the scalability and viable implementation perspective must be considered also. To improve prediction results in the future other features can be added to dataset. Missing values can be eliminated by collecting respective data and more Machine Learning methods can be applied and evaluated.

REFERENCES

[1] Pankaj Ajit Rohit Punnoose, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms".

[2] Predictive Analytics Today website. [Online].
http://www.predictiveanalyticstoday.com/what-is-predictive-analytics

[3]https://trends.google.com/trends/explore?date=2004-01-01%202018-02-01&q=Data%20Science,Big%20Data,Machine%20Learning

[4] H. Jantan, A. R. Hamdan, and Z. A. Othman, “Towards Applying Data Mining Techniques for Talent Managements”, 2009 International Conference on Computer Engineering and Applications, IPCSIT vol.2, Singapore, IACSIT Press, 2011.

[5] V. Nagadevara, V. Srinivasan, and R. Valk, “Establishing a link between employee turnover and withdrawal behaviours: Application of data mining techniques”, Research and Practice in Human Resource Management, 16(2), 81-97, 2008.

[6] W. C. Hong, S. Y. Wei, and Y. F. Chen, “A comparative test of two employee turnover prediction models”, International Journal of Management, 24(4), 808, 2007.

[7] L. K. Marjorie, “Predictive Models of Employee Voluntary Turnover in a North American Professional Sales Force using Data-Mining Analysis”, Texas, A&M University College of Education, 2007

[8] D. Alao and A. B. Adeyemo, “Analyzing employee attrition using decision tree algorithms”, Computing, Information Systems, Development Informatics and Allied Research Journal, 4, 2013

[9] V. V. Saradhi and G. K. Palshikar, “Employee churn prediction”, Expert Systems with Applications, 38(3), 1999-2006, 2011

[10] A. Liaw and M. Wiener, “Classification and regression by randomForest”, R news, 2(3), 18-22, 2002.

[11]] Goel, E. and Abhilasha, E. (2017). Random Forest: A Review. International Journal of Advanced Research in Computer Science and Software Engineering, 7(1), pp.251-257.

[12] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting”, Journal of computer and system sciences, 55(1), 119-139, 1997.

[13] Alao, D., and A. B. Adeyemo. "Analyzing employee attrition using decision tree algorithms." *Computing, Information Systems, Development Informatics and Allied Research Journal* 4 (2013).

[14] A. Amin, F. Rahim, I. Ali, C. Khan, and S. Anwar, A Comparison of Two Oversampling Techniques (SMOTE vs MTDF) for Handling Class Imbalance Problem: A Case Study of Customer Churn Prediction, pp. 215–225. Cham: Springer International Publishing, 2015

[15] L. Breiman, Random forests. Machine Learning, 45(1), 5–32, 2001

[16] Turnover blog. [Online]. http://www.inostix.com/blog/en/4-challenges-with-predictive-employee-turnover-analytics/

[17] example dataset. [Online]. https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/

[18] Pugliese Amedeo, Recker Jan Mertens Willem, Quantitative Data Analysis.

[19] D. Liu, T. R. Mitchell, T. W. Lee, B. C. Holtom, and T. R. Hinkin, “When employees are out of step with coworkers: How job satisfaction trajectory and dispersion influence individual-and unit-level voluntary turnover”, Academy of Management Journal, 55(6), 1360-1380, 2012.

[20] G. King and L. Zeng, “Logistic regression in rare events data”, Political Analysis, 9(2), 137–163, 2001.

[21] Predictive workforce turnover. [Online]. http://www.talentanalytics.com/blog/the-beginners-guide-to-predictive-workforce-analytics/

[22] T. Fawcett, “An introduction to ROC analysis”, Pattern Recognition Letters 27 (8), 861–874, 2006.

[23] A. Liaw and M. Wiener, “Classification and regression by randomForest”, R news, 2(3), 18-22, 2002

[24] L. Breiman, Random forests. Machine Learning, 45(1), 5–32, 2001