SEATTLE PACIFIC UNIVERSITY

SUBJECT: ISM 6359

DATA MINING

TOPIC: Employee Attrition

To build a machine learning model to predict the causes of Employee Attrition in IBM.

Tool: WEKA

Done By

Sheetal Murali

SPU ID no: 900409765

TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| SL NO | CONTENT | PAGE NO. |
| 1. | Introduction | 3 |
| 2. | Business Understanding |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

INTRODUCTION:

IBM is the global leader in business transformation through an open hybrid cloud platform and AI, serving clients in more than 170 countries around the world. Today 47 of the Fortune 50 Companies rely on the IBM Cloud to run their business, and IBM Watson enterprise AI is hard at work in more than 30,000 engagements. IBM has **345,000** total number of employees in 2021**.**

Employees are the backbone of any organization. Its performance is heavily based on the quality of the employees and retaining them. With employee attrition, organizations are faced with a number of challenges:

Expensive in terms of both money and time to train new employees

Loss of experienced employees

Impact on productivity

Impact on profit

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency.

BUSINESS UNDERSTANDING

According to Mactrotrends.com, We can see that IBM has a declining trend in the number of Employees over the years.

Chart, bar chart

Description automatically generated

.

Interactive chart of IBM (IBM) annual worldwide employee count from 2010 to 2022 is shown as above,

IBM total number of employees in 2021 was 345,000, a 0.26% decline from 2020.

IBM total number of employees in 2020 was 345,900, a 9.87% decline from 2019.

IBM total number of employees in 2019 was 383,800, a 0.71% increase from 2018.

IBM total number of employees in 2018 was 381,100, a 4.2% decline from 2017.

Attrition in human resources refers to the gradual loss of employees over time. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture and motivation systems that help the organization retain top employees. Hence, It is important for the organization to uncover the factors that lead to employee attrition and retain the employees.

BUSINESS REASON:

IBM Employee Attrition Prediction

The goal of this project is to identify the causes of employee attrition in IBM through exploratory data analysis and analyze them using different classification models to determine whether an employee is likely to leave. This could significantly improve the HR department's capacity to act quickly to address the issue and stop attrition.

I have chosen WEKA as my tool to conduct this analysis

WEKA is an open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems.

DATA UNDERSTANDING:

The data collected is a high-level overview of what to expect in a data science pipeline and the tools that can be used along the way. It starts from framing the business question, to buiding and deploying a data model. The pipeline is demonstrated through the employee attrition problem.

First, I started with collecting raw data from the [official website](https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem/?mhsrc=ibmsearch_a&mhq=attrition) which had 1470 instances and 35 columns. Hence I decided to explore more and find more data with relevance to the topic from [Kaggle](https://www.kaggle.com/datasets/rushikeshghate/capstone-projectibm-employee-attrition-prediction?select=IBM+HR+Data+new.csv). Here I could find a Capstone Project for Employee Attrition.

Data Structure: There are 37 features that describe each employee, his/her role in the company, his/her level of satisfaction, Education, salary and share options available and his/her influence on the company.

Out of the above, 2 features ( Employee Number, Application ID) are redundant because they contain information that is not relevant to an employee’s decision to leave the company and 2 features (Over 18=yes, Standard Hours=80 hours) with 0 variance. i.e the answer is common to all hence they can be omitted.

This dataset is a supervised learning since the Attrition column is already given. This has 2 classes – Whether the employee is a current employee or he/she has given voluntary resignation.

* Data Structure: 23405 (rows) and 37 (Columns)
* Label: Attrition (Voluntary Resignation, Current Employee)
* Missing Values : 351
* Outliers: Yes
* Data type: Nominal, Numeric and String

DATA PREPARATION:

Once we have determined that we have the right data, then we move on to Data preparation. Sometimes we have too much data or too less data to generate the right model, hence we need the Data Preparation step. A lot of Data Mining projects spend almost 80% of the time on Data preparation. The Data preparation phase includes select the right amount of data for test and train, clean the data to reduce redundancy, deal with missing data, generate new attributes by examining the existing attributes.

For a machine learning algorithm to give acceptable accuracy, it is important to cleanse the data first. This is because the raw data collected from the field contains missing values, irrelevant columns, duplicate entries and so on

**Loading the data:**

When you click on the Explorer button in the Applications selector, On the top, you will see several tabs such as: • Preprocess • Classify • Cluster • Associate • Select Attributes • Visualize.

The data can be loaded from the following sources:

• Local file system

• Web

• Database

1. **Rename:** This is an operator used to Name the attributes as we want. We change the name of attributes in the Parameters after using the Rename operator.
2. **Add ID:** Set an ID to the data set, which acts as an unique Identifier which we do not require hence the algorithm does not use it while calculating the data.
3. **Split Data/Cross Validation**: For my research I have chosen Supervised learning hence, I decided to try both Split data and Cross validation for multiple Algorithms.

**Split data**: 70-30%

**Cross Validation**: 5 Fold and 10 fold for all the algorithms

Graphical user interface, application

Description automatically generated

1. **Set Role**: Here in WEKA, We call the set role as ClassAssigner, Where we set the Label for the data. Alternatively, WEKA uses default last Index as Class.

Graphical user interface, text, application, email

Description automatically generated

1. **Discretize:** In this project I have used Discretize by Binning for different attributes like Age, Monthly Income and Frequency (**PKI Discretize**) for Education, Department. Although there is no option to discretize by User Specification. We can generate attribute by building an equation for the same.

Table

Description automatically generated

1. **Select attributes:** Here we select the necessary Attributes which is used for the analysis. The algorithms only use the selected attributes which thus eliminates the irrelevant information from calculating. we might to do data prep by getting rid of bad data first. The dataset contains 37 attributes of which few can be eliminated. For example, In this project we do not require Employee number, application ID, Over 18 standard hours(same for all), for the calculation, Hence these attributes can be eliminated.
2. **Normalise**: Normalization converts the value of numerical attributes to common scale. Normalizes all numeric values in the given dataset (apart from the class attribute, if set). By default, the resulting values are in [0,1] for the data used to compute the normalization intervals.
3. **Filter outliers:** There are extreme value in the data set which will not fall into the bell curve and thus impact results. It is not recommended to Overfit the regression line because of an outlier as it cannot be used in general business problems. Hence this can be eliminated.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Options without outliers | Options with Outliers |
| Department | Sales, Research & Development, Human Resources | Sales, Research & Development, Human Resources, 1296 |
| Gender | Male, Female | Male, Female, 1, 2 |
| Marital Status | Single, Married, Divorced | Single, Married, Divorced, 1, 2 |
| Overtime | Yes, No | Yes, No, Y |
| Employee source | Referral, Company Website, Indeed, GlassDoor, LinkedIn, Adzuna, Seek, Recruit.net, Jora | Referral, Company Website, Indeed, GlassDoor, LinkedIn, Adzuna, Seek, Recruit.net, Jora, Test, 15, 1 |

1. **NominatoBinary**: Converts all nominal attributes into binary numeric attributes. Example: Overtime: yes=0, no=1 and

Gender: Female=0,Male=1

1. **Rename Nominal Values**: In Weka we do not have Nominal to numeric, instead we use Rename Nominal Attribute. We can easily convert those nominal values using the "Rename Nominal Value" filter by adding values. Decision tree, RF accept Nominal values but other more powerful algorithms like SVM, Deep Learning etc accept only numeric values.

**Graphical user interface, text, application, email

Description automatically generated**

Hence I converted the following attributes during Pre-Processed stage to generate accurate results:

1. Age: '(-inf-21.5]':1, '(21.5-26.5]':2, '(26.5-27.5]':3, '(27.5-29.5]':4, '(29.5-33.5]':5, '(33.5-57.5]':6, '(57.5-58.5]':7, '(58.5-inf)':8
2. Marital Status- Single:1, Divorced:2, Married:3
3. BusinessTravel - Travel\_Rarely:1, Travel\_Frequently:2, Non-Travel:3
4. Department: Sales:1, Research & Development:2, Human Resources:3
5. EducationField: Life Sciences:1, Technical Degree:2, Medical:3, Marketing:4, Other:5, Human Resources:6
6. JobRole: Sales Executive:1, Manager:2, Research Director: 3, Sales Representative:4, Laboratory Technician:5, Research Scientist: 6, Manufacturing Director:7, Healthcare Representative:8, Human Resources:9
7. Employee source- Referral:1, Company Website:2, Indeed:3, GlassDoor:4, LinkedIn:5, Adzuna:6, Seek:7, Recruit.net:8, Jora:9
8. JobSatisfaction- Low:1, Medium:2, High:3, Very High:4
9. PerformanceRating: Low:1, Good:2, Excellent:3, Outstanding:4
10. RelationshipSatisfaction: Low:1, Medium:2, High:3, Very High:4
11. WorkLifeBalance: Bad:1, Good:2, Better:3, Best:4
12. JobInvolvement: Low:1, Medium:2, High:3, Very High:4
13. EnvironmentSatisfaction: Low:1, Medium:2, High:3, Very High:4
14. Education: Below College:1, College:2, Bachelor:3, Master:4, Doctor:5
15. **Add Expression**: This works similar to generate attribute, where An instance filter that creates a new attribute by applying a mathematical expression to existing attributes. In this model, I used the following expression to create JobSatEval

Where Job Satisfaction Evaluation attribute is calculated by finding the average of similar attributes that affect job satisfaction which are:

JobSatisfaction, PerformanceRating, RelationshipSatisfaction, WorkLifeBalance, JobInvolvement, EnvironmentSatisfaction

(a7+a10+a13+a20+a21+a25)/6

Graphical user interface, text, application, email

Description automatically generated

1. Principle Component Attribute: After generating attributes, I still had 27 attributes which is a huge amount of data to process, Hence to derive accurate results I used PCA (Feature Selection). This performs a principal components analysis and transformation of the data.

Graphical user interface, text, application, email

Description automatically generated

The Dimentionality reduction has ranked attributes based on their individual Evaluations. In this project the highest rank of an attribute had varience of 2.111. Hence I chose the first 13 attributes which is almost the half of the threshold of highest varience.

Text

Description automatically generated

DATA MINING ALGORITHMS:

For this analysis, I used the Classification method. The data collected is based on supevised learning. This dataset contains both input and output parameters, it is said to be labeled. In other words, the correct answer has already been assigned to the data. Here, I created a labeled data set which is the Attrition.

Classification It is a commonly used technique for categorizing data points. It involves using algorithms that can be easily modified to improve the data quality. The primary goal of classification is to connect a variable of interest with Attrition.

The algorithm establishes the link between the variables for prediction. The algorithm used for classification in data mining in WEKA is called the classifier, and observations made through the same are called the instances.

Firstly, I split data into Training set and Test set: 70-30 split

Secondly I used Cross Validation for better results: 5 folds and 10 folds

I have used the following algorithms to compare the performances:

1. Decision Tree (Random Tree in WEKA)
2. Random Forest
3. KNN
4. Support Vector Machine ( SMO in WEKA)
5. Ensemble method (Voting)
6. Neural Network (Multilayer Perceptron in WEKA)

DECISION TREE

|  |  |  |
| --- | --- | --- |
| Accuracy | Split Data  (70%-30%) | Cross-Validation  (10 fold) |
| Correctly Classified Instances | 97.8078 % | 98.868 % |
| ROC Curve |  |  |
| Confusion Matrix  Classified |  |  |

RANDOM FOREST – 1000 Trees

|  |  |  |
| --- | --- | --- |
| Accuracy | Split Data  (70%-30%) | Cross-Validation  (10 fold) |
| Correctly Classified Instances | 98.7758 % | 99.4192 % |
| ROC Curve |  |  |
| Confusion Matrix |  |  |

SMO : Sequential Minimal Optimization. Svm configuration in WEKA

|  |  |  |
| --- | --- | --- |
| Accuracy | Split Data  (70%-30%) | Cross-Validation  (10 fold) |
| Correctly Classified Instances | 84.2135 % | 84.1696 % |
| ROC Curve |  |  |
| Confusion Matrix |  |  |

K NEAREST NEIGHBOUR

|  |  |  |
| --- | --- | --- |
| Accuracy | Split Data  (70%-30%) | Cross-Validation  (10 fold) |
| Correctly Classified Instances | 99.3594 % | 99.4876 % |
| ROC Curve |  | A picture containing graphical user interface  Description automatically generated |
| Confusion Matrix |  |  |

Neural Network

Chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
| Accuracy | Split Data  (70%-30%) | Cross-Validation  (5 fold) |
| Correctly Classified Instances | 85.3808 % | 84.9639 % |
| ROC Curve |  |  |
| Confusion Matrix |  |  |

Voting - Ensemble Method

Test Mode: Cross Validation – 10 folds

Correctly Classified Instances = 23258 which gave 99.321% accuracy

Incorrectly Classified Instances =159 with 0.679 %

Total Number of Instances = 23417

Text, letter

Description automatically generated

ROC Curve:

Graphical user interface, application

Description automatically generated

ANALYSIS OF THE OUTPUT

1. From the above analysis, We found that cross validation has better accuracy than splitting data.

Decision tree - 97.807%

Random Forest – 99.41%

SVM- 84.16%

KNN – 99.48%

Neural Network – 84.96%

1. Although the highest result is from Knn with K value 20, I have performed an Ensemble method to derive the best result for this model.
2. The Area under Curve is the highest for Knn Model which is 99.97%
3. From the Voting Algorithm I could determine that the Correctly Classified Instances have 99.321% accuracy and the Area under curve is 98%. This technique created multiple models and then combined them to produce improved results.
4. Every model makes a prediction (votes) for each test instance and the final output prediction is the one that receives more than half of the votes, this is the final prediction
5. After building the final model and determining the accuracy, this can be used by the organization to implement in determining whether a given employee is likely to resign or stay in the company.

REFERENCES

<https://www.ibm.com/us-en?ar=1>

[IBM: Number of Employees 2010-2022 | IBM | MacroTrends](https://www.macrotrends.net/stocks/charts/IBM/ibm/number-of-employees)

<https://www.youtube.com/@WekaMOOC>

Dataset source:

The data is made available publicly under the following license agreements:

[https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem//?mhsrc=ibmsearch\_a&mhq=attrition](https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem/?mhsrc=ibmsearch_a&mhq=attrition)

<https://github.com/IBM/employee-attrition-aif360/blob/master/data/emp_attrition.csv>

Further, I found the dataset used to build this model on Kaggle.com

<https://www.kaggle.com/datasets/rushikeshghate/capstone-projectibm-employee-attrition-prediction/code>