PROJECT REPORT

HR ANALYTICS PROJECT-

UNDERSTANDING THE ATTRITION IN HR

SUBMITTED BY

NAME : NATASHA PODDAR

BATCH: 1843

PROBLEM DEFINITION

Artificial Intelligence is one of the most unique and innovative technologies that is most extensively adopted by different companies in making various strategic decisions relating to organizational and operational aspects, People Management, Finance and other departments of the company. In today’s world both subjective and objective data analysis is essential is driving critical decisions. The main motive is to find out and understand what are the various factors that influence an employee’s attrition in a company. The aim is to understand and identify the major causes that contribute to an employee’s decision of staying in the company and once we have an understanding of this, we should be able to predict whether an employee will resign or not.

A company invests a lot in an employee in terms of training and development programs, productivity loss in the beginning of his employment, raise and increment, incentives etc. and when an employee decides to leave the company an added cost for looking out for his replacement, the productivity loss which takes place in case of a vacancy left, the extra cost incurred on training the new employee etc. becomes a massive cost for the company. To overcome this issue, we have this study where our data set consists of about 35 different independent features and attributes and close to 1500 rows where we try and understand that what are the most important features that stimulate an employee to continue working in the same company and to be able to create a good scoring classification model which will help us in predict an employee’s attrition and will assist the company in making timely decisions to avoid extra cost.

DATA ANALYSIS

If we talk first on the descriptive Data Analysis, where out of 1470 total attrition responses, 1233 (84%) were negative, they decided to stay in the company where as 237 were positive (close to 16%).

The breakup of the various departments are the following:

* R & D team recorded the highest figures of resigned employees (56%)
* Sales Department recorded an attrition of 20%
* Human Resource recorded an attrition of 19%

Non Graphical Univariate Data Analysis:

We observed the mean, median, min, max, std deviation, etc. figures of all int/float columns and had a fair understanding of skewness and outliers presence in these columns

Graphical Univariate Data Analysis:

Count-plots, Histograms and dist-plots of every column was produced and analyzed

Graphical Bi-variate Data Analysis:

The relationship between the target and all other features were established, the different relationship between independent features were also analyzed

Graphical Multi-variate Data Analysis:

Box plots were drawn which helped in analyzing the different relationship between 3 features

EDA CONCUDING REMARKS

The various insights drawn from the analysis are the following :

* Monetary Income is the most essential contributor of employee attrition
* With an increase in Salaries, there is a decrease in attrition
* Employees with fewer years of experience are more likely to

leave the company

* With an increase in distance from home there is a progressive increase in the attrition rate too
* Younger employees within the age group of 18-23 are more likely to resign from the company
* Salary increase (“PercentSalaryHike”) and performance evaluation (“PerformanceRating”):

employees who receive an increase in salary tend to be more productive;

* Salary (“MonthlyIncome”) and employment level (“JobLevel”) -

employees with greater seniority

generally tend to earn more;

* Spent years of work in the company (“YearsAtCompany”), spent years with the current manager
* (“YearsWithCurrManager”) and spent years in the current role (“YearsInCurrentRole”): this

would seem to highlight the absence of professional growth

* Results show that job involvement of the employee in the processes or tasks of the company is the

most inﬂuencing parameters for his attrition

* Females have shown a high attrition even though their monthly income is higher than males
* R & D and Sales dept have shown same % salary hike where are hr has shown a lower % across both the genders
* Job involvement is uniform and so is attrition across both the genders
* Work life balance is same across both the depts across both the genders
* Both males and females belong to similar age group and females have a higher attrition
* Attrition is almost equal across all depts and life science, medical, marketing and technical degree have a similar daily rate
* Monthly income of managers is higher across all job roles and has the highest attrition
* All males and females have a similar marital status whereas the hourly rate for divorced females are less
* Overtime is slightly more in females as compared to males and they have a higher monthly income too
* Marketing and technical degree have similar number of years since last promotion and attrition is higher in marketing and highest in others
* Divorced employees have high attrition and they have worked in more number of companies
* Environment satisfaction is fairly similar across all dept except for divorced employees belonging to research scientist fields and the satisfaction is the least in single employees in research director field
* Married employees have a higher monthly income and they travel frequently the most
* Attrition is max across married employees and they have a higher number of years with current managers
* Employees are similarly from all the different job roles and sales executive have highest attrition and they have worked in max companies too
* Marketing field has employees who have received promotion many years ago with lesser attrition as compared to employees of other group
* Employees in HR dept have shown the highest income and they have the highest overtime figures across all fields
* Males and females located within 2-12 km from the organization are less likely to resign as compared to those who stay further away
* When the no of year's since last promotion is more than 2 attrition is zero and vice versa
* Monthly rate is consistent when the no of years in the company range from 0-10 , then it scatters gradually
* As years in current role increase, years in the company also increases
* As age increases, number of years in the company also increase
* As total working years increase, years in current role increase too
* Monthly rate ranges between roughly 2000-25000 between employees whose work experience range from 1-26 years and then it slowly becomes discreet
* Stock option level 1 has the last count whereas attrition is higher across stock option level 0

PREPROCESSING PIPELINE

First and foremost, we identified if there were any duplicate values or not in the data set.

The data was thoroughly cleaned, no null values were found in the data set. Unique and count of each column was observed.

Graphical and Non-Graphical, Univariate, Bivariate and Multivariate Analysis were conducted where we analyzed the relationship between different variables.

Min ,50th, mean, median ,75th and max figures were observed and worked on

The following categorical columns were encoded:

* Attrition
* Business Travel
* Education Field
* Department
* Education
* Job Role
* Marital status
* Over 18
* Overtime

The two columns listed below were deleted from the Data Set

* Over 18
* Employee Count
* Standard working hours

Correlation was checked with respect to all independent features and along with target.

Skewness was checked both visually and numerically and was removed by Power Transform except on categorical data

Outliers were checked both visually and numerically and was removed timely with the help of ZSCORE

Columns which had outliers are the following :

* NumCompaniesWorked
* YearsSinceLastPromotion
* TrainingTimesLastYear
* StockOptionLevel
* YearsWithCurrManager
* MonthlyIncome
* TotalWorkingYears
* YearsInCurrentRole
* Attrition

Data Loss was roughly 5.6% which was within our acceptable range hence we moved ahead with the newer data set

The data set was then separated into X & Y, where VIF scores were being checked and accordingly the data set was being amended.

We balanced the data set through SMOTE then reduced the number of columns using Principal Component Analysis.

Post this we started training and testing the data set respectively and worked upon

Various Model libraries along with metrics and were called

Our final data set XPCA & Y1 for model testing and training :

BUILDING MACHINE LEARNING MODELS

* X, containing all independent variables;
* Y containing the dependent variable, i.e., “Attrition”;

Once the pre-processing was complete, both X and Y were established and we chose the following models where our algorithms were chosen

* Logistic Regression
* Support Vector Machines – Classifier
* K Nearest Neighbor – Classifier
* Decision Tree -Classifier
* Random Forest Classifier
* Gradient Boosting Classifier

20% of the Data set taken for testing where as 80% was conducted on training,

In order to allow the model to learn the relationships hidden in the data; the train-set contains

1176 observations

Test set contained the remaining 30%. This information was dedicated to the test and validation

phase in order to evaluate the general performance of the model and to calculate errors between

predicted and actual results; the test-set contains 294 observations.

The Accuracy received on the models are mentioned below :

* Logistic Regression (79.74%)
* Support Vector Machines – Classifier (91.16%)
* K Nearest Neighbor – Classifier (84.91%)
* Decision Tree -Classifier (83.83%)
* Random Forest Classifier (92.24%)
* Gradient Boosting Classifier (85.86%)

AUC-ROC curve was drawn for each model to understand the TRUE & FALSE POSITIVE RATES

Cross-validation:

We adopted this technique to prevent over-ﬁtting problems and to simplify the model. The training-set was randomly divided into TEN parts (k)—one was used as a validation-set and the other k-1s as training-sets, repeating the procedure k times. In each of the iterations, a different part was taken as the validation-set and ﬁnally the average prediction error was obtained by assessing the average errors in the k iterations performed on each k-validation set

We took out the difference between each model accuracy score and CV mean score where we found that

Random Forest Classifier, not only had the least difference between cv mean score and model accuracy score, it also has the highest accuracy compared to all other models, hence we hyper tuned this model

Post hyper tuning it the accuracy score had improved to 93.75 from 92.24 wrt to the cv mean score, hence we selected & saved this model

CONCLUDING REMARKS

This phase evaluated the qualities of the adopted models. The results of the decisions made in the prediction phase were collected, for each algorithm, in the relative “confusion matrix”. This is a matrix where the values predicted by the classiﬁer are shown in the columns and the real values of each instance of the test-set are shown in rows. To proceed with the performance evaluation, we used the confusion matrix to derive a series of fundamental metrics to quantitatively express the efﬁciency of each algorithm: recording accuracy, precision, recall, speciﬁcity and F1-score

* What are the key indicators that signal that an employee will leave the company?
* What is the probability that an employee will leave the company?

To this aim, we applied some machine learning techniques in order to identify the factors that may contribute to an employee leaving the company and, above all, to predict the likelihood of individual

employees leaving the company. First, we assess statistically the data and then we classiﬁed them.

The dataset was processed, dividing it into the training phase and the test phase, guaranteeing the same distribution of the target variable

We selected various classiﬁcation algorithms and, for each of them, we carried out the training and validation phases. To evaluate the algorithm’s performance, the predicted results were collected and fed into the respective confusion matrices. From these it was possible to calculate the basic metrics necessary for an overall evaluation

(precision, recall, accuracy, f1 score, ROC curve, AUC, etc.) and to identify the most suitable classiﬁer to predict whether an employee was likely to leave the company or not.

The algorithm that produced the best results for the available dataset was the Random Forest classiﬁer and the same was selected and saved for further predictions:

Results obtained by the proposed automatic predictor demonstrate that the main attrition variables are monthly income, age, overtime,

distance from home. The results obtained from the data analysis represent a starting point in the development of increasingly efﬁcient employee attrition classiﬁers.