

DATA SCIENCE EXTERNSHIP **PROJECT - SMARTBRIDGE**

Problem statement:

**Predicting Employee Attrition Using
Random Forest with IBM Cloud.**

TEAM MEMBERS:

MEERA KRISHNAKUMAR

V RAGHAV ANAND

PARVATHY AJ

S. SAI SREELEKHA

P. SRI SNEHA KARTHIKA

Introduction:

Employee resignations are a reality for any business. However, if the situation isn't handled properly, key staff members' departures can lead to a downturn in productivity. The organization may have to employ new people and train them on the tool that is being used, which is time consuming. Most organizations are interested in knowing which of their employees are at the risk of leaving. This paper discusses the application of the Random Forest algorithm as a method of predicting employee attrition. This is done by using data from Kaggle and treating the problem as a classification task. We conclude the project by deploying the model using a Flask framework.

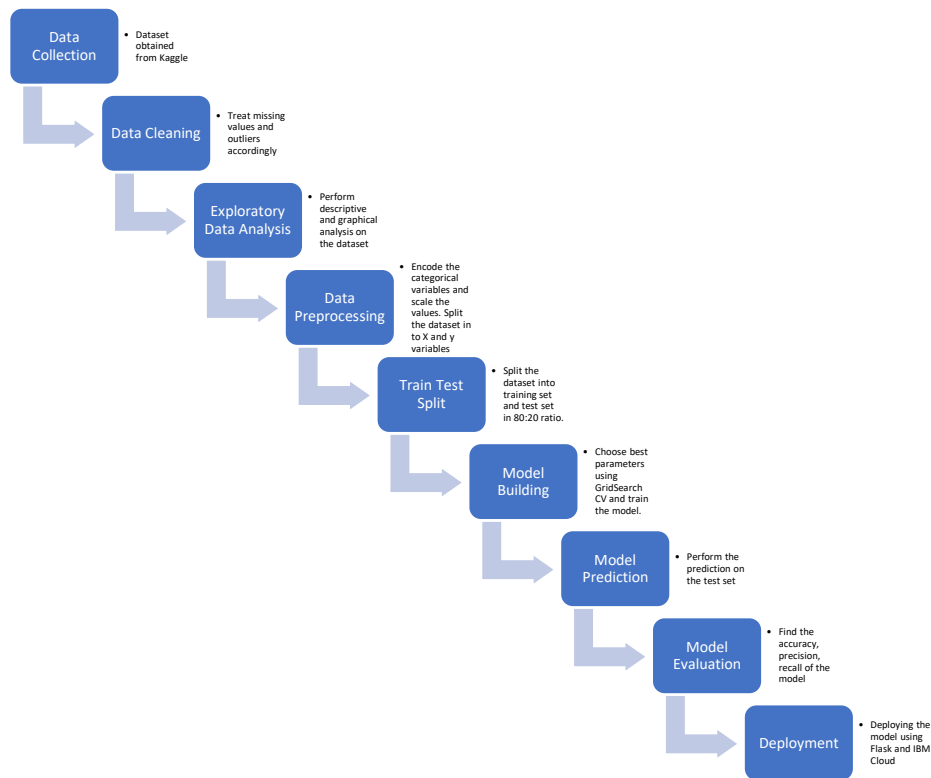
Literature Survey:

Employee attrition refers to the gradual loss of employees over time. Most literature on employee attrition categorizes it as either voluntary or involuntary. Involuntary attrition is thought of as the mistake of the employee, and refers to the organization firing the employee for various reasons. Voluntary attrition is when the employee leaves the organization by his own will. This paper focuses on voluntary attrition. A meta-analytic review of voluntary attrition found that the strongest predictors of voluntary attrition included age, pay, and job satisfaction. Other studies showed that several other features, such as working conditions, job satisfaction, and growth potential also contributed to voluntary attrition. Organizations try to prevent employee attrition by using machine learning algorithms to predict the risk of an employee leaving, and then take pro-active steps for preventing such an incident.

Related Work:

Human resource management (HRM) has been shown to be effective in working scenarios, production and management, and identifying correlations with productivity by a number of researches [2,3]. In fact, the findings show that HRM's impact on productivity has a beneficial impact on a company's capital growth and intensity [4]. Most studies [5,6] focus on analysing and monitoring customers and their behaviour, but often ignore a company's most valuable asset: its workers. Employee attrition has been the subject of numerous studies. Employee demographics and job-related features, such as compensation and length of employment relationship, are the factors that most effect employee attrition, according to previous study [7]. Another study [8] looked at the effects of demographic factors and employee absenteeism on employee attrition. The authors of [9] concentrated solely on work-related issues. In [10], the authors compared the performance of a Nave Bayes classifier and the J48 decision tree algorithm in predicting the likelihood of an employee leaving the company. For each algorithm, two techniques were tested: tenfold cross-validation and percentage split 70. Using tenfold cross-validation, J48 had an accuracy of 82.4 percent and an inaccurate classification of 17.6 percent, whereas percentage split 70 had an accuracy of 82.7 percent and an incorrect classification of 17.3 percent. Using tenfold cross-validation, the Nave Bayes classifier achieved an accuracy of 78.8% and an inaccurate classification of 21.2 percent, whereas percentage split 70 achieved an accuracy of 81 percent and an incorrect classification of 19 percent. Authors in [11] investigated the use of Logistic Regression to predict employee turnover and found that it had an accuracy of 85% and a false negative rate of 14%.

Flow chart:



Methodology:

Following are the steps that were performed:

Data Acquiring:

The dataset used has been obtained from Kaggle. The dataset consists of 14999 entries, each entry describing a particular employee. The various attributes are:

- Satisfaction Level of the employee
- How long it has been since their last evaluation, in years
- The number of projects they have been part of.
- Their average monthly hours
- The number of years spent in company.
- Have they had a work accident or not.
- Have they received a promotion in the last 5 years.
- Their department
- Their salary category.

The dependent variable is whether the employee left or not.

Data Cleaning:

The dataset we obtained was clean. There were no missing values. The attribute number of years in company had outliers but we decided not to treat them, since they were not arbitrarily high values, the maximum value in this column was 10 years, which is normal for any employee to spend in a company.

Exploratory Data Analysis:

The Department column had 10 different categories out of which three were, technical, IT, support. Since all three of them might mean the same thing in a company, we decided to club these 3 categories together as 'technical'.

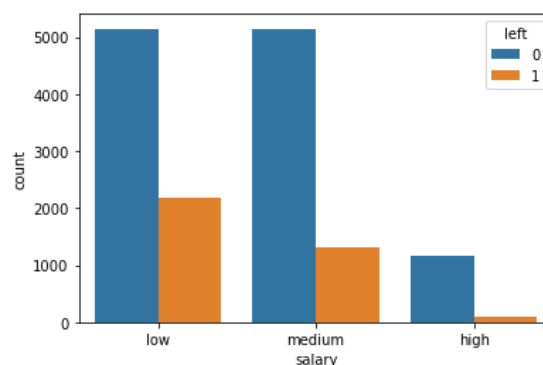
Next, we made analysis on all the numerical attributes based on our dependent variable, left.

	satisfaction_level	last_evaluation	number_project	average_monthly_hours	time_spend_company	Work_accident	promotion_last_5years
left							
0	0.666810	0.715473	3.786664	199.060203	3.380032	0.175009	0.026251
1	0.440098	0.718113	3.855503	207.419210	3.876505	0.047326	0.005321

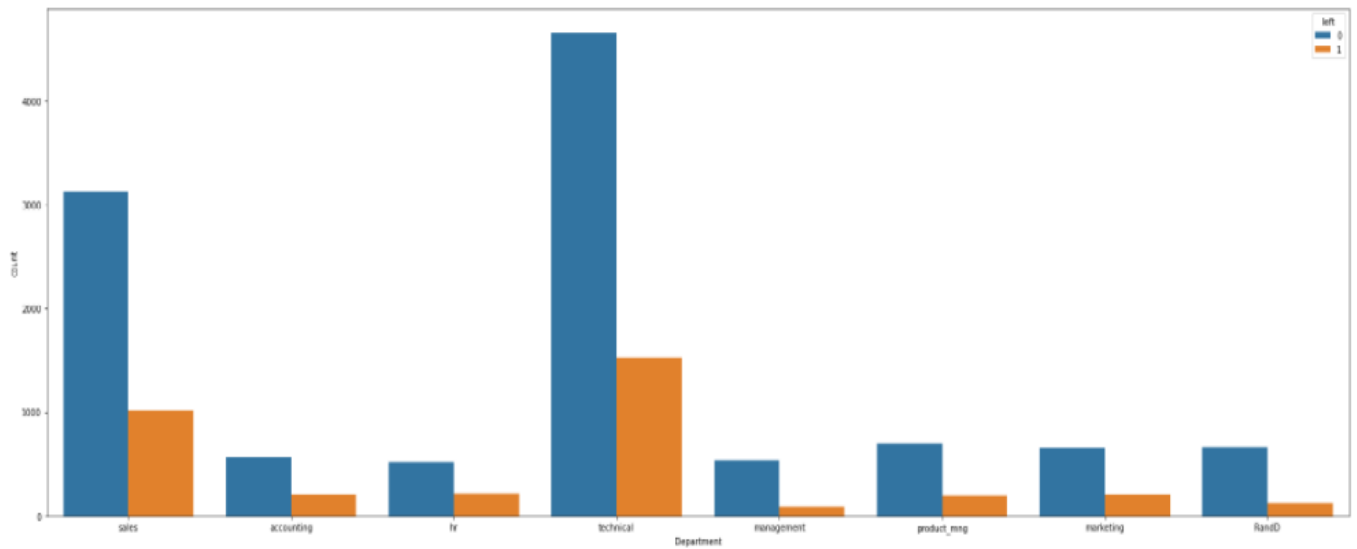
The following observations were made:

- Employees who do not leave the company have a higher satisfaction level
- Employees with higher average monthly hours are more likely to leave.
- An employee who received a promotion in the last 5 years is less likely to leave.
- If the employee had a workplace accident, he/she prefers to stay in the company

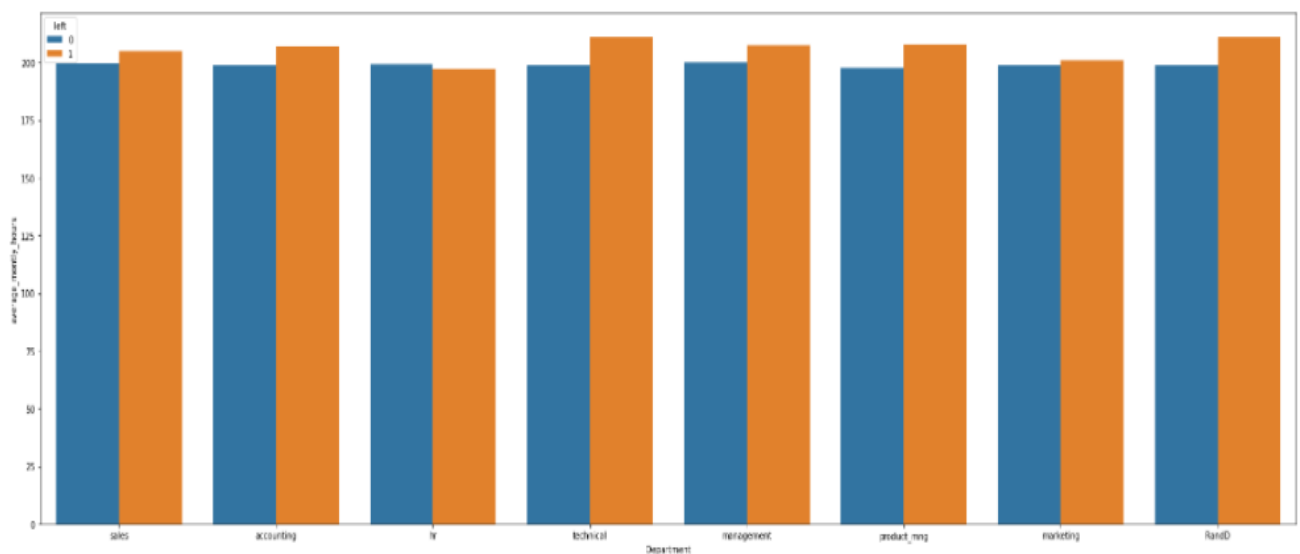
An analysis on Salary vs count of employees based on whether they left or not showed that employees with 'low' salary are more likely to leave and employees with high salaries are likely to stay.



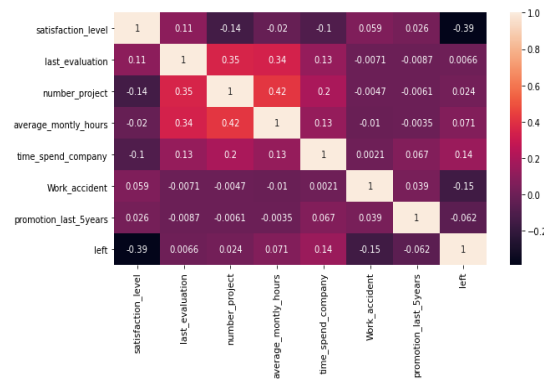
Department of the employee does not seem to have much impact on the leaving decision of the employee.



It was observed that there was not much difference in the monthly working hours of the employees across various departments.



The correlation heat map of the dataset showed that the time since last evaluation attribute had very less dependence on the dependent variable, left and hence was removed as part of our analysis.



Data Pre-processing:

The dataset was divided into X and y for independent variable and dependent variable respectively.

Since the dataset had two categorical columns – Department and Salary, they were encoded using One Hot Encoding.

Since, we had a wide range of values, we decided to scale all the attributes using Standard Scaler. Standard Scaler takes any value and transforms it using the following formula:

$$X' = \frac{X - \mu}{\sigma}$$

Train-Test Split

The model was split into training and test set in the ratio of 80:20.

Category	Size
X Train	11999
X Test	3000
Y Train	11999
Y Test	3000

Model Building

Random Forest Algorithm was used for the model building. The best parameters for the model were obtained using GridSearch CV, through which we obtained the best parameter as 9 for number of trees in the forest(n_estimators) and 9 as the maximum depth of the forest.

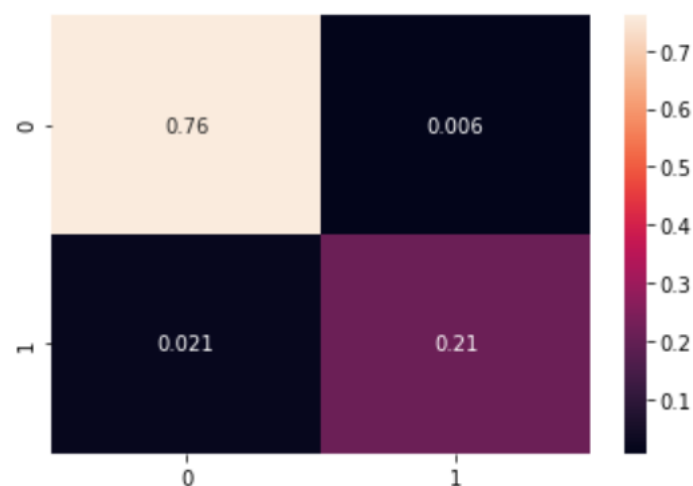
GridSearch CV took the range from 1-13 for n_estimators and max_depth as 2 to 10. For all combinations the dataset was each time spit into groups of 5 and the best model was obtained.

Model Prediction and Evaluation

After training the model, it was tested on the test set to obtain a testing accuracy of 97.33%. The precision and recall were also calculated as per the below table

	precision	recall	f1-score	support
0	0.97	0.99	0.98	2299
1	0.97	0.91	0.94	701
accuracy			0.97	3000
macro avg	0.97	0.95	0.96	3000
weighted avg	0.97	0.97	0.97	3000

The confusion obtained is shown below, the values represent the percentage of values.

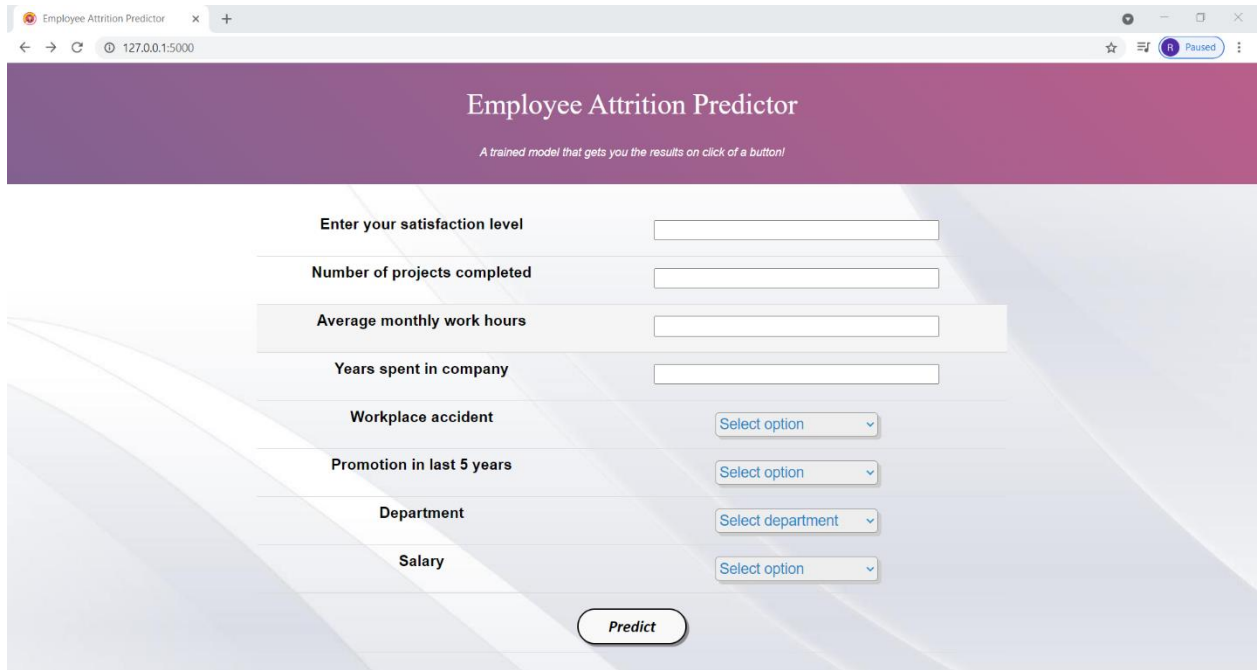


Deployment using Flask and IBM Cloud

An appropriate UI was created and the model was deployed using Flask and integrated with IBM Cloud using the scoring endpoint.

Results:

USER INTERFACE:

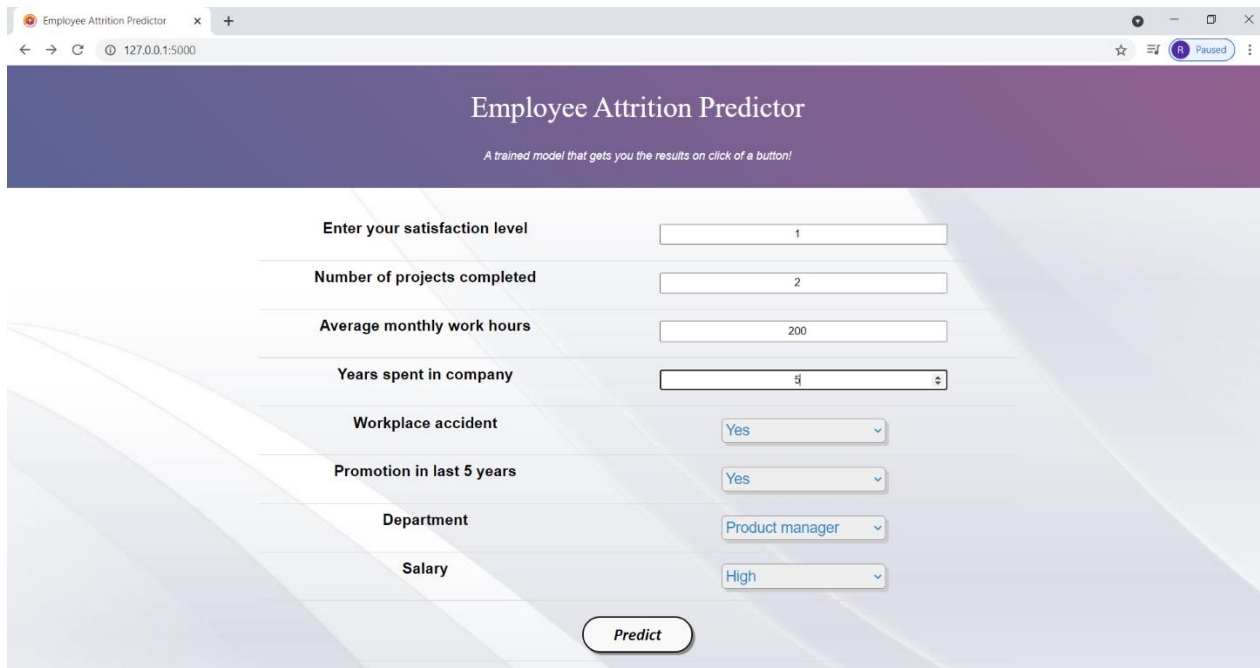


The screenshot shows a web browser window with the title "Employee Attrition Predictor". The URL bar shows "127.0.0.1:5000". The page has a purple header with the title and a subtitle "A trained model that gets you the results on click of a button!". Below the header, there is a form with the following fields:

- Enter your satisfaction level:
- Number of projects completed:
- Average monthly work hours:
- Years spent in company:
- Workplace accident:
- Promotion in last 5 years:
- Department:
- Salary:

At the bottom of the form is a "Predict" button.

FILLING THE FORM IN SUCH A WAY THAT THE PREDICTION SAYS THE EMPLOYEE WON'T LEAVE THE COMPANY:

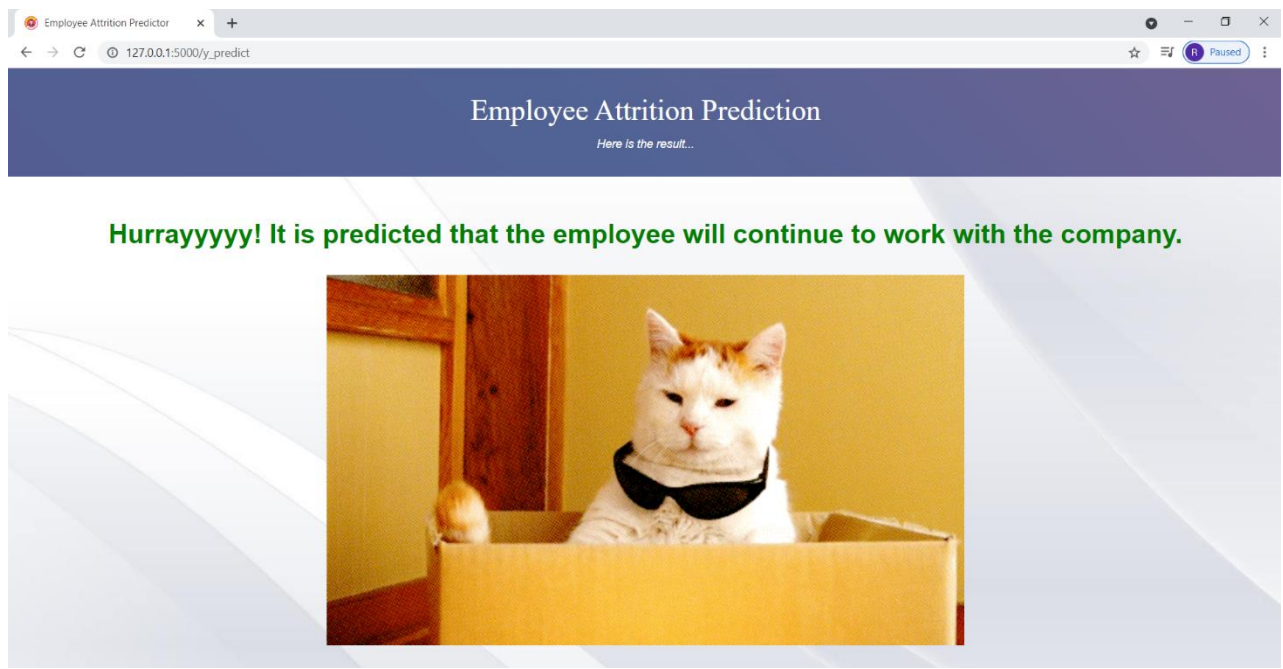


The screenshot shows the same web browser window as the previous one, but with the form fields filled out. The values entered are:

- Enter your satisfaction level: 1
- Number of projects completed: 2
- Average monthly work hours: 200
- Years spent in company: 5
- Workplace accident: Yes
- Promotion in last 5 years: Yes
- Department: Product manager
- Salary: High

The "Predict" button is still visible at the bottom of the form.

PREDICTION OBTAINED AFTER CLICKING THE PREDICT BUTTON:



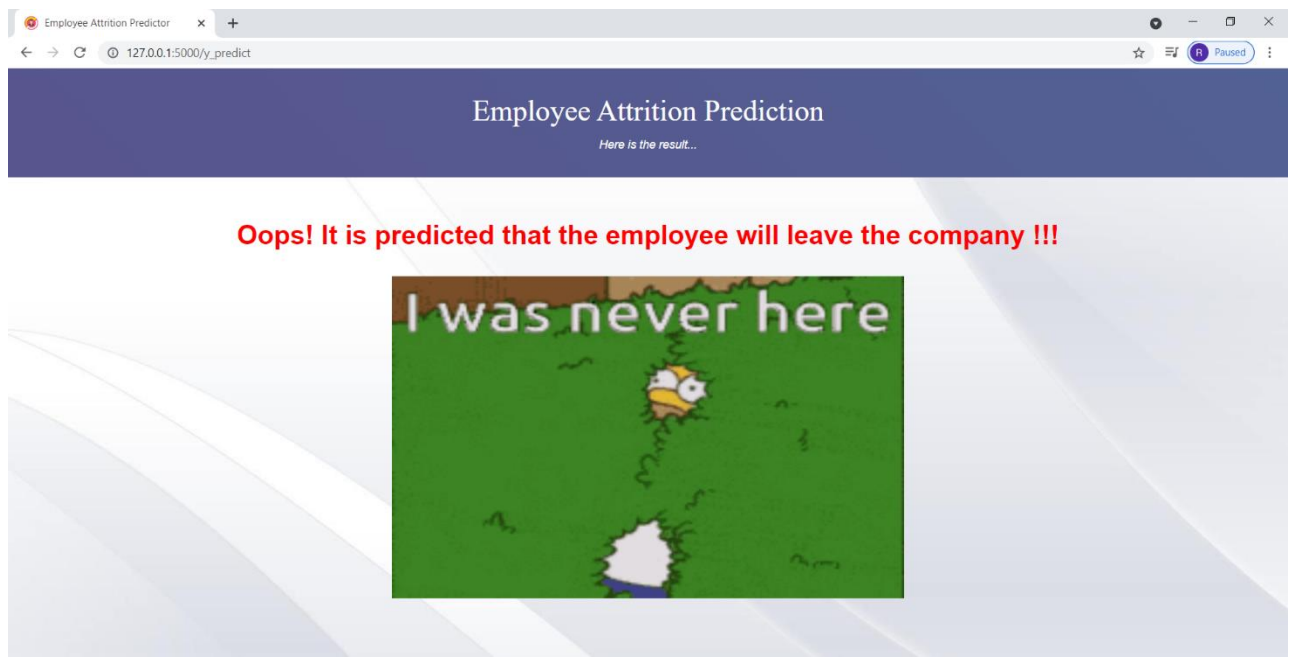
FILLING THE FORM IN SUCH A WAY THAT THE PREDICTION SAYS THE EMPLOYEE WILL LEAVE THE COMPANY:

The screenshot shows the "Employee Attrition Predictor" web application in its input form state. The header is dark blue with the title "Employee Attrition Predictor" and the subtitle "A trained model that gets you the results on click of a button!". The form contains the following fields:

Field	Value
Enter your satisfaction level	0.38
Number of projects completed	2
Average monthly work hours	157
Years spent in company	3
Workplace accident	Yes
Promotion in last 5 years	No
Department	Technical
Salary	Low

At the bottom of the form is a "Predict" button.

PREDICTION OBTAINED AFTER CLICKING THE PREDICT BUTTON:



Advantages:

- Allows the company to assess where they have to work for employee retaining.
- Gives the rough estimate to the company of the costs it will incur depending on the kind of employee.
- The Company can apply appropriate strategies to retain the employee.

Disadvantages:

- Though the attributes can help in predicting the retention, but the reason an employee leaves can be out of the scope of the attributes which we are using.
- There might have been a transfer and hence, the employee left the company which is not clearly indicated in the dataset.

Conclusions:

Long-term success and health of the company are based on the retention of skilled employees that helps in meeting the expectations of client and increasing the productivity of the business. With evidence of issues in the IT sector, it is important for any company's management to take corrective actions for reducing the employee turnover. For this purpose, research can be conducted to assess the satisfaction level of employees and potential reasons behind the high attrition rate. It facilitates to understand whether employees are satisfied or not so as to provide the quick solution of the issues which are being faced by the business. As a result, it's much easier to lower attrition rates and, as a result, boost total productivity and profitability. Furthermore, predicting employee turnover will assist the organisation in

identifying qualified candidates for future vacancies. Simultaneously, it aids management in providing appropriate encouragement to employees based on their level of contentment or intent to leave the organisation. Furthermore, the organisation will be able to change its expansion or productivity plans or targets. As a result, predicting the attrition rate is critical to ensuring the business's continued growth and development. This will also assist in preserving the company's greater rate of return. The main objective of this research was to use machine learning models to predict employee attrition based on their features. This will give company management signs supported by machine learning tools. As a result, this will help management to act faster to reduce the likelihood of talented employees leaving their company. We have used the Random Forest Algorithm and tried to model this classification problem and have achieved a 97.3% accuracy. We have also deployed this model using Flask. Future work can include predicting the kind of bonus/ strategy the company should formulate to retain the employees.

References:

- [1] Dataset - <https://www.kaggle.com/giripujar/hr-analytics>
- [2] Marchington, M.;Wilkinson, A.; Donnelly, R.; Kynighou, A. Human Resource Management atWork; Kogan Page Publishers: London, UK, 2016.
- [3] Van Reenen, J. Human resource management and productivity. In Handbook of Labor Economics; Elsevier: Amsterdam, The Netherlands, 2011.
- [4] Deepak, K.D.; Guthrie, J.; Wright, P. Human Resource Management and Labor Productivity: Does Industry Matter? Acad. Manag. J. 2005, 48, 135–145.
- [5] Gordini, N.; Veglio, V. Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. Ind. Mark. Manag. 2016, 62, 100–107. [CrossRef]
- [6] Keramati, A.; Jafari-Marandi, R.; Aliannejadi, M.; Ahmadian, I.; Mozaffari, M.; Abbasi, U. Improved churn prediction in telecommunication industry using data mining techniques. Appl. Soft Comput. 2014, 24, 994–1012. [CrossRef]
- [7] Alao, D.; Adeyemo, A. Analyzing employee attrition using decision tree algorithms. Comput. Inf. Syst. Dev. Inf. Allied Res. J. 2013, 4, 17–28.
- [8] Nagadevara, V. Early Prediction of Employee Attrition in Software Companies- Application of Data Mining Techniques. Res. Pract. Hum. Resour. Manag. 2008, 16, 2020–2032.
- [9] Rombaut, E.; Guerry, M.A. Predicting voluntary turnover through Human Resources database analysis. Manag. Res. Rev. 2018, 41, 96–112. [CrossRef]
- [10] Usha, P.; Balaji, N. Analysing Employee attrition using machine learning. Karpagam J. Comput. Sci. 2019, 13, 277–282.
- [11] Ponnuru, S.; Merugumala, G.; Padigala, S.; Vanga, R.; Kantapalli, B. Employee Attrition Prediction using Logistic Regression. Int. J. Res. Appl. Sci. Eng. Technol. 2020, 8, 2871–2875. [CrossRef]