Employee Attrition

Grace Sigalla
Drexel University
gas83@drexel.edu

Sreerekha Rajendran
Drexel University
sr3547@drexel.edu

Abhijit Purru
Drexel University
ap3792@drexel.edu

***Abstract*—Employers may face a lot of difficulty as a result of employee attrition, also known as voluntary employee turnover. An essential component of efficient human resource management is the ability to anticipate employee attrition.**

This study examines how to forecast employee attrition in businesses using Pyspark. In order to find trends and factors that contribute to employee attrition, the study involved the examination of a dataset of employee records. Machine learning models were built using Pyspark to predict employee turnover based on features like employee satisfaction, salary, department, promotion, and time spent on the job. The results of this study can assist organizations in taking proactive measures to keep their employees and efficiently manage attrition.

Keywords: attrition, pyspark, mllib, logistic regression, random forest, decision tree, svm

I. INTRODUCTION

Employees voluntarily leaving a company, or employee attrition, is a major issue for many organizations. Significant financial losses, decreased productivity, and lowered morale among the remaining employees can all be brought about by the departure of valuable employees. Employee attrition is thus a significant concern for employers and human resource managers, who must work proactively to retain staff and reduce attrition.

Machine learning (ML) has become an effective technique for forecasting employee turnover in recent years. Large datasets of employee records can be analyzed by ML models to find trends and causes of attrition. Organizations can take proactive measures to keep employees and manage attrition more successfully by using ML to forecast employee turnover.

In this study, we investigate the usage of Pyspark, a Python-based framework for distributed computing, to forecast staff attrition.

The rest of this paper is structured as follows. We review the data and its source. Then we outline our methodology for assessing the dataset which is pre-processing, create the machine learning models and evaluate their performances. Thereafter, we examine the implications of our findings and offer suggestions for businesses looking to effectively manage employee attrition.

II. DATASET DESCRIPTION

The dataset chosen for this task is based off an HR employee record available on kaggle. It has 14999 records and 10 columns. The features tell us about the satisfaction level of the employees, their last evaluation, number of projects worked on, average monthly hours, the time they spent in the company, if or not they had any work accident, their promotion history for the last five years, their department and lastly salary. We have a column named 'left' which shows if the employee had left the organization or not, which is what we are trying to predict through this project and hence it is our label.

```
df.printSchema()
```

```
root
 |-- satisfaction_level: double (nullable = true)
 |-- last_evaluation: double (nullable = true)
 |-- number_project: integer (nullable = true)
 |-- average_montly_hours: integer (nullable = true)
 |-- time_spend_company: integer (nullable = true)
 |-- Work_accident: integer (nullable = true)
 |-- left: integer (nullable = true)
 |-- promotion_last_5years: integer (nullable = true)
 |-- Department: string (nullable = true)
 |-- salary: string (nullable = true)
```

Fig. 1. Dataset Schema

Below are the number of unique values in each column.

```
Unique values for column 'satisfaction_level' are: 92
Unique values for column 'last_evaluation' are: 65
Unique values for column 'number_project' are: 6
Unique values for column 'average_montly_hours' are: 215
Unique values for column 'time_spend_company' are: 8
Unique values for column 'Work_accident' are: 2
Unique values for column 'left' are: 2
Unique values for column 'promotion_last_5years' are: 2
Unique values for column 'Department' are: 10
Unique values for column 'salary' are: 3
```

Fig. 2. Unique values

III. RELATED WORK

In recent years, numerous research have been carried out to address the problem of staff attrition. Here are some of the key works of research in this field.

- Lama Alaskar and Martin Crane's[1] "Employee Turnover Prediction Using Machine Learning Algorithms" (2019) - In this study, employee turnover in a business was predicted using machine learning techniques. To forecast employee turnover, the authors utilized a variety of classification techniques, such as de-

cision trees, logistic regression, and support vector machines (SVM). In their research, they discovered that SVM outperformed other classification algorithms at predicting employee turnover.

- Rohit and Ajit's, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms(2016)"[2] - In these studies, the authors predicted employee turnover using data mining techniques and ML. To forecast staff churn, they used a variety of classification techniques, such as decision trees, random forests, xgboost, and SVM. In both the studies, the researchers discovered that when it came to forecasting employee turnover, XGBoost outperformed other categorization systems.

The potential of machine learning approaches to forecast staff attrition is generally highlighted by these studies. Our study expands on such prior studies by predicting staff attrition in a corporate environment using PySpark.

IV. EXPLORING DATA

A. Exploratory Data Analysis

We performed exploratory data analysis (EDA) to learn more about the factors influencing employee retention. We investigated any trends or patterns that would be helpful in forecasting attrition using our EDA, as well as the relationships between these variables and employee attrition. Our goal is to create predictive models that can assist businesses in lowering staff turnover, boosting productivity, and raising employee happiness by combining our EDA with machine learning algorithms.

We started with looking at the distribution of the no. of employees who left v/s who stayed.

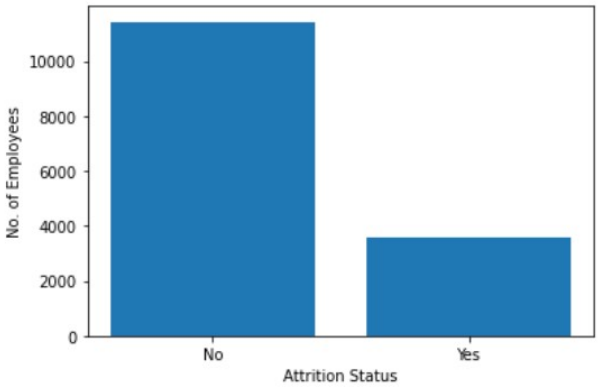

Fig. 3. Bar plot showing the two classes

The following plots tells us that the attrition rates are higher among low income compared to high and medium income group; and that more than half of the employees left the company within the first 3 years of employment.

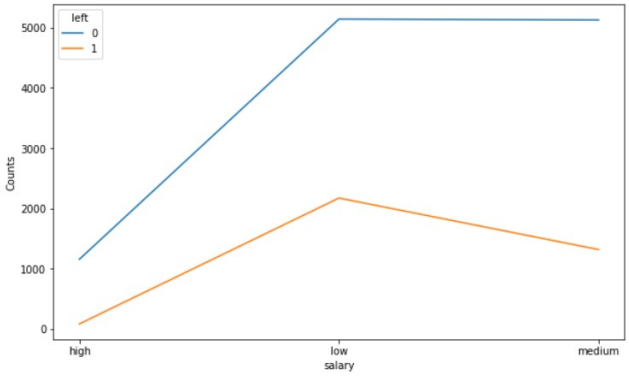

Fig. 4. Attrition rate with Salary

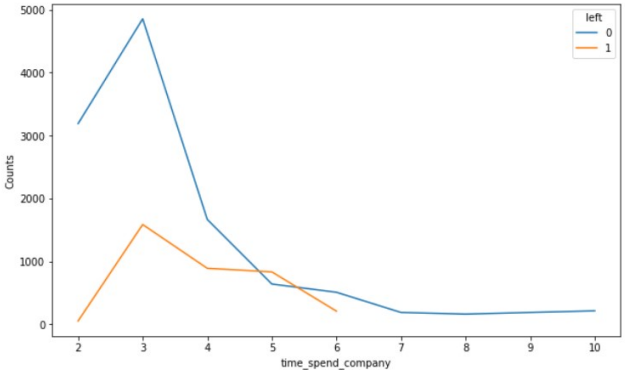

Fig. 5. Attrition rate with the time spend in the company by the employees

B. Pre-processing before Modeling

Pre-processing is a crucial stage in Pyspark modeling since it helps to convert the data into a format that machine learning algorithms can easily understand. StringIndexer, which converts each categorical variable into a distinct numerical value, is a popular pre-processing method for categorical variables.

The next step is to use a One-Hot Encoder to transform the numerical values into a binary vector. Each distinct numerical value is converted into a binary vector via the One-Hot Encoder. Because it enables the machine learning algorithms to treat each category as an independent feature, this strategy is advantageous.

After utilizing StringIndexer and One-Hot Encoder to transform the categorical variables, all the features are combined into a single vector using the Vector Assembler. The machine learning algorithms can then use this vector as input.

In conclusion, StringIndexer, One-Hot Encoder, and Vector Assembler are essential components of Pyspark's preprocessing process because they convert input into a format that is appropriate for machine learning methods. This method enhances the models' precision and offers more insightful analyses of the data.

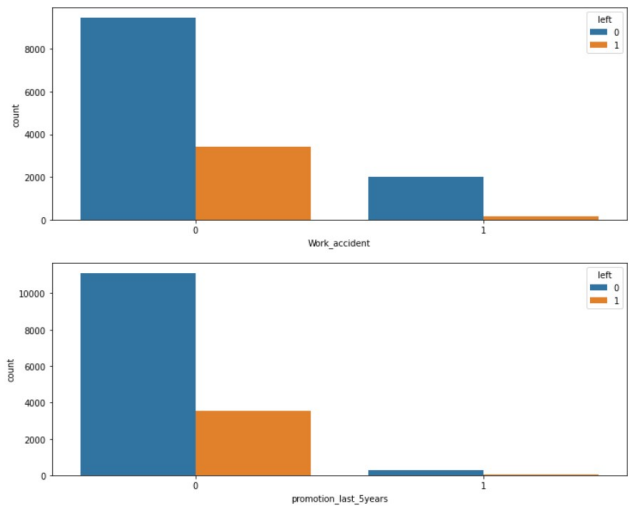

Fig. 6. Attrition rates w.r.t work accidents and promotion in the last 5 years

V. RESULTS AND DISCUSSION

A. Logistic Regression

Logistic regression is a classification algorithm used to classify data into one of two possible outcomes in supervised machine learning[3]. The accuracy from the logistic regression model for our dataset is 0.78.

AUC	0.82
Accuracy Score	0.78
Precision	0.75

Fig. 7. Metrics on test set predictions on LR model

B. Random Forest

Random forest is an ensemble learning machine learning algorithm that builds many decision trees and combines them to get a more accurate and stable predictions[3]. Random Forest relies on bagging, or bootstrap aggregation, which randomly selects a subset of data points to train each decision tree. The accuracy from the random forest model of our dataset is 0.97.

AUC	0.99
Accuracy Score	0.97
Precision	0.97

Fig. 8. Metrics on test set predictions on RF model

C. Decision Tree

Decision Tree is a decision support tool that utilizes tree-like graph or model decisions to show a statistical probability of an outcome[3]. The accuracy from the decision tree model of our dataset is 0.96.

AUC	0.95
Accuracy Score	0.968
Precision	0.968

Fig. 9. Metrics on test set predictions on DT model

D. Linear Support Vector Machine

Linear support vector machine (SVM) is an algorithm that can be used for classification and regression tasks by maximizing the margin between two classes[3]. The accuracy from SVM model of our dataset is 0.78.

AUC	0.80
Accuracy Score	0.78
Precision	0.75

Fig. 10. Metrics on test set predictions on SVM model

E. Naive Bayes

Naive Bayes is a supervised machine learning algorithm that uses the Bayes' theorem with the "naïve" assumption of independence of each pair of features[3]. The accuracy from Naïve Bayes model of our dataset is 0.76.

AUC	0.57
Accuracy Score	0.76
Precision	0.72

Fig. 11. Metrics on test set predictions on NB model

F. Gradient Boosted Tree

Gradient Boosted Tree (GBT) combines multiple decision trees in order to produce a more powerful model. It uses boosting algorithm to train each decision tree[3]. It is particularly useful for handling complex, non-linear relationships between the features and the target variable. However, GBT can also be prone to overfitting if the number of trees is too high, or if the learning rate is too low. The accuracy from Gradient Boost Tree model of our dataset is 0.96.

AUC	0.98
Accuracy Score	0.965
Precision	0.965

Fig. 12. Metrics on test set predictions on GBT model

From all six models we implemented, Random Forest had the highest accuracy of 97%, followed by Decision Tree and Gradient Boost at 96%.

Comparison of the models as seen on the graph below:

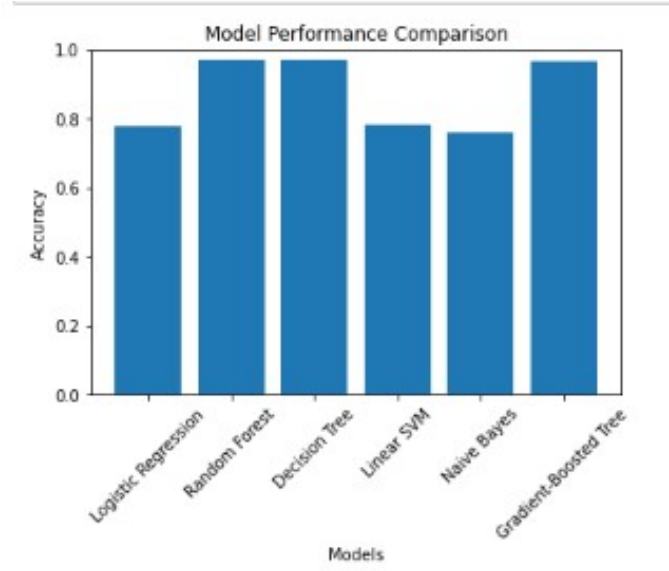

Fig. 13. Model comparison

VI. CONCLUSIONS

From our analysis in this project, we observed that employee satisfaction level, the time they spent in the company, the number of projects they worked on, and their average monthly hours are some of the most important features that help in the attrition prediction. One would expect salary to be the most important predictor, but to our surprise, in this case it is not one of the most important features.

The application of data science in people analytics may help businesses target high-risk and high-performing individuals for retention initiatives, and it can even help them make evidence-based decisions about incentive payments and equity grant distributions. In today's fast-changing and highly competitive world, this knowledge is vital.

VII. FUTURE WORK

- Predicting Employee Performance: Data scientists can use employee data to create predictive models that forecast employee performance and assess their future potential.
- Understanding Employee Engagement: Data scientists can use employee data to measure employee engagement and provide insights on how to improve it. This could include looking at factors such as job satisfaction, team dynamics, and the impact of company policies on employees.
- Analyzing Workflows: Data scientists can use employee data to analyze how workflows are structured and how they could be improved. This could involve examining ways to reduce redundancies, improve communication, or streamline processes.
- Creating Employee Retention Strategies: Data scientists can use employee data to create strategies for retaining high-performing employees. This could include examining factors such as salary, benefits, and workplace culture.
- Improving Recruiting Efforts: Data scientists can use employee data to analyze the effectiveness of the company's recruiting efforts. This could involve examining what types of candidates are being hired, how long it takes for them to be hired, and how well they fit into the company culture.

VIII. APPENDIX

The code file, dataset and readme for this project are uploaded to GitHub repository: https://github.com/GraceSigalla/pyspark_employee_attrition_project

REFERENCES

[1] Alaskar, L., Crane, M. and Alduailij, M., Employee turnover prediction using machine learning, SpringerLink. Springer International Publishing. Available at: https://link.springer.com/chapter/10.1007/978-3-030-36365-9_25.

[2] Prediction of employee turnover in organizations using machine learning (2016). Available at: https://www.researchgate.net/publication/308043155_Prediction_of_Employee_Turnover_in_Organizations_using_Machine_Learning_Algorithms.

[3] Arora, L. (2020) Building Machine Learning Pipelines using Pyspark, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2019/11/build-machine-learning-pipelines-pyspark/ (Accessed: March 3, 2023).