ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 12, December 2024

Employee Churn Prediction Using Python

[1] Gracy Samson, [2] Medini Shirpurkar, [3] Vaibhavi Naik, [4] Jaideep Singh Chopra, [5] Suraj Khandare

[1] [2] [3] [4] BE IT, FCRIT, Vashi, Maharashtra
[5] Professor, FCRIT, Vashi, Maharashtra

Corresponding Author Email: [1] gracysamson7@gmail.com, [2] smedini88@gmail.com, [3] naikvaibhavim@gmail.com, [4] jaideepsinghchopra029@gmail.com, [5] suraj.khandare@fcrit.ac.in

***Abstract*— In an era where workforce turnover poses a significant challenge for organizations, our Python-based project addresses workforce turnover by developing an employee churn prediction system. Utilizing historical employee data, including job satisfaction and performance metrics, we employ Python's data analysis and machine learning capabilities for model construction. Through rigorous testing, our project identifies the most effective predictive model for employee churn, providing valuable insights to enhance workforce management strategies. Aimed at reducing churn and improving stability, this data-driven solution offers organizations a tool to revolutionize human resources practices, fostering increased employee satisfaction and loyalty.**

***Index Terms*— Employee churn prediction, python, machine learning, data analytics, talent retention.**

I. INTRODUCTION

In an era characterized by dynamic shifts in the workforce landscape, the ability to predict and mitigate employee churn has become a paramount concern for organizations. Recognizing the pressing need for proactive talent retention strategies, our project endeavors to introduce an innovative approach to address this challenge. By harnessing the power of Python and machine learning, we aim to develop and assess predictive models that can foresee employee churn, thus empowering businesses to take informed measures to retain their valuable staff.

This project's inception stems from the pivotal role that employee churn prediction plays in ensuring not only the organizational efficiency but also the well-being of the workforce. Aligning with the broader trend of leveraging technology in human resources management, it is imperative to leverage data-driven solutions to enhance employee retention.

Our initiative seeks to establish a more accurate and effective mechanism for predicting employee churn, transcending the limitations associated with conventional methods. The primary objective is to identify and evaluate the most suitable machine learning algorithms for this purpose, a challenge that underpins the project's core.

The project's scope encompasses the collection and analysis of historical employee data, encompassing a variety of influential factors, such as job satisfaction, performance metrics, compensation, and more. It involves data preprocessing, feature engineering, and the application of machine learning algorithms to construct predictive models. Through rigorous testing and optimization, we aim to determine the most effective model for predicting employee churn.

In a corporate world where talent retention is paramount, our Python-powered project aspires to equip organizations with data-driven insights to reduce employee churn and enhance workforce stability. By embracing Python's analytical capabilities, we aim to provide organizations with a transformative tool to revolutionize their human resources practices, ultimately leading to a more stable and content workforce.

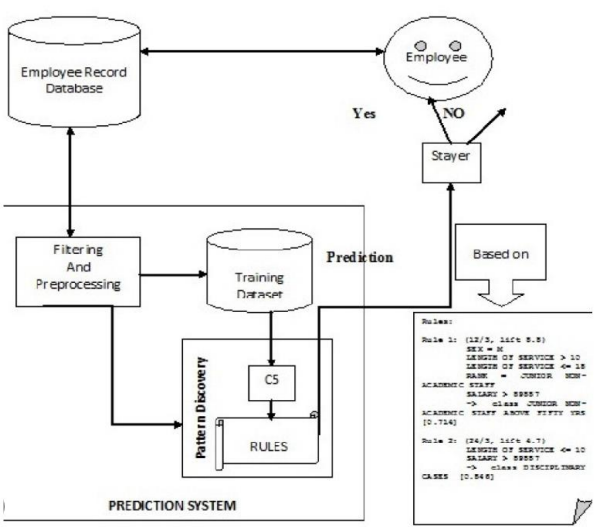


Fig. 1. Approach

II. RELATED WORK

A. Employee Churn Prediction using Logistic Regression and Support Vector Machine (2021)

Presents the development of machine learning models for predicting employee churn using two popular classification algorithms, Logistic Regression and Support Vector Machine (SVM). These models were trained on the IBM HR employee dataset from Kaggle and subsequently fine-tuned to enhance their performance. Various performance metrics, including precision, recall, confusion matrix, AUC, and ROC curve, were utilized to assess model performance. The paper highlights the use of grid search (GridSearchCV) for

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 12, December 2024

optimizing Logistic Regression, focusing on parameters like C, kernel, and gamma for the SVM model. These findings contribute to the advancement of employee churn prediction methods using machine learning and provide insights for HR and organizational decision-makers.

B. Employee Turnover Prediction with Machine Learning (2021)

Offers a comprehensive exploration of supervised machine learning methods for predicting employee turnover within organizations. The study involves extensive numerical experiments conducted on both real and simulated human resources datasets, representing entities of varying sizes, including small, medium, and large employee populations. A wide array of machine learning techniques is employed, encompassing decision trees, random forests, gradient boosting trees, extreme gradient boosting, logistic regression, support vector machines, neural networks, linear discriminant analysis, Naïve Bayes, and K-nearest neighbor methods. The paper goes beyond merely showcasing these methods and rigorously evaluates their performance for employee turnover prediction using robust statistical methodologies. This research significantly advances the field of employee turnover prediction by offering a comprehensive and methodical assessment of various machine learning techniques.

C. Employee Churn Prediction in Healthcare Industry using Supervised Machine Learning (2022)

Delves into the critical issue of employee churn within the healthcare sector, a concern that can significantly impact an organization's productivity and operational costs. A variety of prediction models, ranging from traditional classifiers like Random Forest, Support Vector Machine, and Logistic Regression, to advanced techniques such as XGBoost and Artificial Neural Network (ANN), are implemented. The results of this investigation reveal the superiority of the Support Vector Machine (SVM) model, achieving an impressive recall rate of 94.8% and a ROC-AUC accuracy of 91.1%. Furthermore, the study employs model-agnostic interpretability methods to identify key contributors to employee churn in the healthcare sector, highlighting factors such as wellness programs, employment rates, vacation days, and sick days as strong indicators. This research not only addresses a critical issue within the healthcare industry but also showcases the effectiveness of machine learning in predicting employee churn, offering actionable insights for retention strategies and aiding in the long-term sustainability of healthcare organizations.

D. Predictive Modeling of Employee Churn Analysis for IoT- Enabled Software Industry (2022)

Delves into the critical realm of employee churn analytics in the context of the rapidly evolving IoT-enabled software industry. Employee turnover can be a costly issue for organizations as they seek to replace departing talent with equally skilled professionals, impacting productivity and incurring additional costs. To address this challenge, the paper explores the use of machine learning techniques that leverage sensor-based data, a relatively unexplored avenue in the corporate sector. The research entails several distinct steps: data selection and preprocessing, the application of four filter-based methods to analyze features, and the selection of the most influential variables from these methods. Remarkably, the proposed technique showcases a remarkable 98% accuracy rate in predicting future churners within IoT-enabled corporate sector organizations. This research contributes significantly by leveraging IoT technology for more precise employee churn prediction, offering valuable insights for HR and management in mitigating the challenges of workforce attrition within the software industry.

III. PROPOSED METHODOLOGY

Adopts a structured approach to analyze and predict employee attrition. It begins with collecting and preprocessing employee data to ensure its accuracy and usability. Key features influencing churn are identified through feature engineering, forming the basis for predictive modeling. Machine learning algorithms, such as logistic regression and random forests, are trained on this data to build effective models.

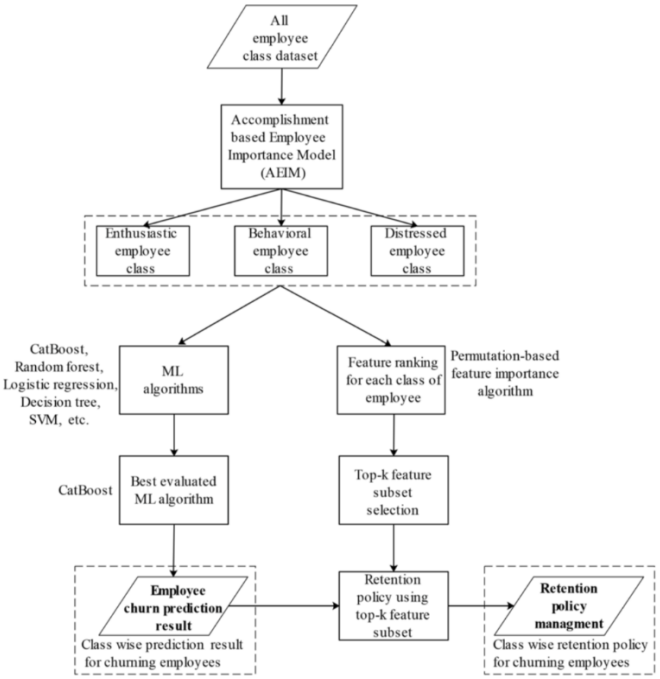


Fig. 2. System Architecture

The architectural design comprises several interconnected components aimed at achieving the project's objectives seamlessly. At the core of this architecture are the data collection and preprocessing components. The data collection phase involves gathering historical employee data,

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 12, December 2024

including a multitude of relevant factors such as job satisfaction, performance metrics, compensation, and more. Subsequently, the data undergoes comprehensive preprocessing to ensure its quality and suitability for analysis. This includes addressing missing values, scaling features, and encoding categorical variables. Feature engineering, where meaningful attributes are extracted from the pre-processed data to identify the most influential factors contributing to employee churn. These features are then utilized in the selection and training of predictive models. Various machine learning algorithms, including decision trees, logistic regression, and random forests, each meticulously trained on the prepared dataset to effectively predict employee churn. Incorporates a model evaluation stage where the performance of the trained models is rigorously assessed using diverse metrics, such as accuracy, precision, recall, and the F1-score. This evaluation process aids in identifying the most effective predictive model, which is further fine-tuned through hyperparameter optimization to achieve optimal performance. An integral aspect of the architecture is the in-depth analysis component, which seeks to uncover the key factors contributing to employee turnover. This analysis provides insights into the complex relationships between these factors and employee churn, offering valuable information for organizational decision-making.

To enhance the system's practical applicability, the proposed methodology incorporates real-time prediction capabilities through an integrated deployment framework. This involves the utilization of a Flask-based web interface, enabling HR professionals to input employee data and instantly receive churn predictions. The integration ensures user-friendly interaction while maintaining robust backend processing. Moreover, the methodology emphasizes scalability and adaptability by implementing cloud-based storage and computational resources, facilitating seamless handling of large-scale datasets and diverse organizational requirements. Advanced techniques, such as cross-validation and ensemble learning methods, are employed to further improve the robustness and generalization of the predictive models. Finally, the framework includes a feedback loop that continuously monitors the system's performance, allowing for iterative improvements and updates based on new organizational trends and data patterns. This comprehensive approach ensures that the predictive system remains relevant and effective in dynamically evolving workplace environments.

IV. RESULTS

In this project, aim is not only to analyze HR Employee Attrition Data but also to develop a robust framework for predicting employee attrition using advanced machine learning techniques. We meticulously preprocess the data, ensuring its suitability for training various models. Through extensive visualization, we gain valuable insights into the dataset's characteristics, aiding in feature selection and model interpretation.

The project stands out for its comprehensive approach, encompassing a wide array of machine learning algorithms ranging from traditional logistic regression and decision trees to more sophisticated neural networks and ensemble methods. By comparing and contrasting these models, we provide a nuanced understanding of their strengths and weaknesses in predicting employee attrition.

Moreover, project's functionality is extended by integrating it with the Flask framework, enabling seamless interaction with users through a user-friendly interface. With user authentication in place, our application ensures data privacy and security, making it suitable for deployment in real-world scenarios. By bridging the gap between machine learning research and practical application, our project contributes significantly to the field of human resources management and predictive analytics.

The predictive models employed in this study offer valuable insights into employee churn dynamics. Logistic Regression, delving into factors like job satisfaction and performance metrics, estimates the likelihood of turnover. Decision Trees visually represent decision-making processes, highlighting critical determinants such as low job satisfaction. Leveraging Random Forest, an ensemble of Decision Trees, enhances accuracy by discerning intricate patterns in employee data. Neural Networks, inspired by human brain functionality, unveil nuanced relationships between variables, providing a comprehensive understanding of attrition drivers. Ensemble methods like Stacking Classifier and Voting Classifier amalgamate individual model predictions, contributing to a sophisticated churn prediction system. These techniques collectively empower organizations with a nuanced and accurate tool for effective workforce management, emphasizing the significance of ensemble methods and individual model contributions in achieving robust predictions for mitigating employee churn.

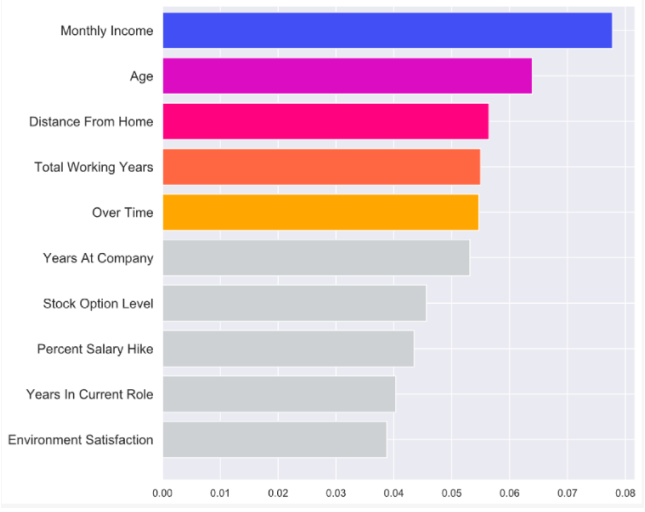
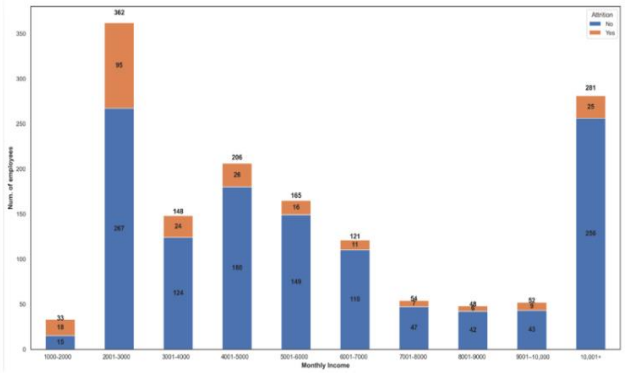


Fig. 3. Feature Importance

ISSN (Online) 2394-2320

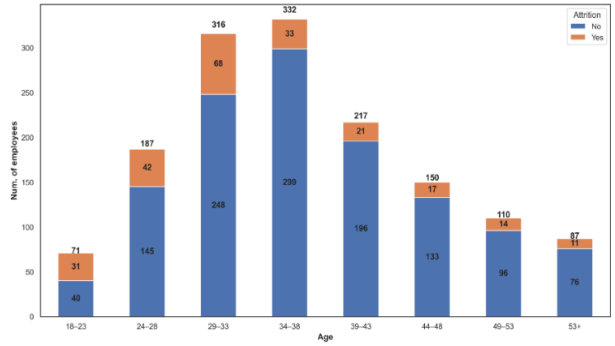
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 12, December 2024



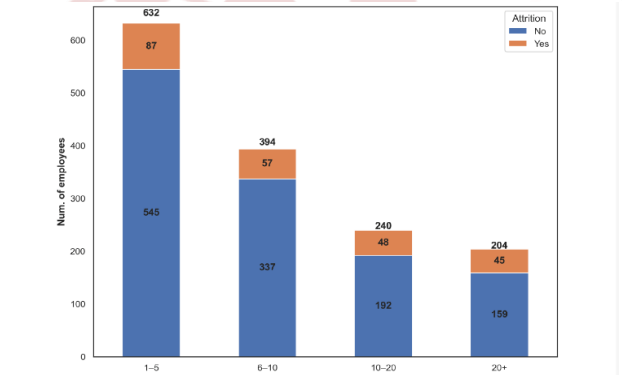
	MonthlyIncomeBins	Num. of employees	Num. of employees with attrition	% attrition in the cluster	% of the total attrition
0	1000-2000	33	18	54.5	7.6
1	2001-3000	362	95	26.2	40.1
2	3001-4000	148	24	16.2	10.1
3	4001-5000	206	26	12.6	11.0
4	5001-6000	165	16	9.7	6.8
5	6001-7000	121	11	9.1	4.6
6	7001-8000	54	7	13.0	3.0
7	8001-9000	48	6	12.5	2.5
8	9001-10,000	52	9	17.3	3.8
9	10001+	281	25	8.9	10.5

Fig. 4. Distribution of attrition by monthly income



	AgeBins	Num. of employees	Num. of employees with attrition	% attrition in the cluster	% of the total attrition
0	18–23	71	31	43.7	13.1
1	24–28	187	42	22.5	17.7
2	29–33	316	68	21.5	28.7
3	34–38	332	33	9.9	13.9
4	39–43	217	21	9.7	8.9
5	44–48	150	17	11.3	7.2
6	49–53	110	14	12.7	5.9
7	53+	87	11	12.6	4.6

Fig. 5. Distribution of attrition by age



	DistanceFromHomeBins	Num. of employees	Num. of employees with attrition	% attrition in the cluster	% of the total attrition
0	1–5	632	87	13.8	36.7
1	6–10	394	57	14.5	24.1
2	10–20	240	48	20.0	20.3
3	20+	204	45	22.1	19.0

Fig. 6. Distribution of attrition by distance from home

V. CONCLUSION

To sum up, the implementation of predictive models in our "Employee Churn Prediction with Python" project represents a significant advancement in the domain of workforce analytics and organizational management. By utilizing a combination of machine learning algorithms, such as logistic regression and random forests, along with a meticulously pre-processed dataset, this project has successfully demonstrated a robust approach to predicting employee attrition. Through the integration of feature engineering and hyperparameter optimization, we have enhanced the accuracy and reliability of the models, ensuring that the predictions are both actionable and insightful.

The project not only addresses the immediate challenge of predicting employee churn but also delves into uncovering the underlying factors that contribute to it. These insights empower organizations to take proactive measures to retain talent by improving job satisfaction, addressing compensation issues, and fostering a positive work environment. By bridging the information gap between employers and employees, our methodology provides a comprehensive framework for data-driven decision-making that enhances organizational stability and reduces turnover rates. Moreover, the project's emphasis on rigorous model evaluation, including metrics like accuracy, precision, recall, and F1-score, ensures the reliability and validity of the predictions. The analysis of key factors influencing churn further positions this work as a critical tool for strategic planning and operational efficiency in the corporate landscape.

In conclusion, this project epitomizes a forward-thinking approach to addressing one of the most pressing challenges in human resource management. It establishes a benchmark for the application of predictive analytics to workforce retention, combining technical rigor with practical relevance. By offering a powerful toolset for understanding and mitigating employee attrition, this research not only contributes to academic and industry discourse but also lays a strong foundation for fostering long-term organizational success in a highly competitive business environment.

REFERENCES

[1] https://www.analyticsvidhya.com/blog/2021/11/employee-attrition-prediction-a-comprehensive-guide/

[2] https://www.mdpi.com/2073-431X/9/4/86

[3] https://github.com/dintellect/EmployeeChurn_Analysis_and_Prediction

[4] https://ieeexplore.ieee.org/abstract/document/8541242/figures#figures

[5] https://towardsdatascience.com/cracking-the-employee-attrition-problem-with-machine-learning-6ee751ec4aae