EMPLOYEE RETENTION PREDICTION

CS19643 – FOUNDATIONS OF MACHINE LEARNING

Submitted by

MADHAN SHANKAR G **(2116220701149)**

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

BONAFIDE CERTIFICATE

Certified that this Project titled **"EMPLOYEE RETENTION PREDICTION"** is the bonafide work of **"MADHAN SHANKAR G (2116220701149)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. V.Auxilia Osvin Nancy, M.Tech., Ph.D.,
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai-602 105.

Submitted to Mini Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

ABSTRACT

With the increasing volatility of financial markets and the strategic importance of gold as a hedge against inflation and economic uncertainty, accurately predicting gold prices has become a critical task for investors, financial analysts, and policymakers. This study introduces a machine learning-based framework designed to forecast gold prices using a diverse set of economic and financial indicators. These features include historical gold prices, crude oil prices, interest rates, inflation data, stock market indices, and currency exchange rates, among others. By leveraging publicly available financial datasets, the project evaluates and compares the effectiveness of various machine learning algorithms such as Linear Regression, Random Forest, Support Vector Machines (SVM), and Gradient Boosting.

The core objective of this research is to identify which model provides the most accurate and reliable predictions of gold prices, measured on daily or monthly time scales. In addition to comparing model performance, the study also analyzes the relative importance of different economic indicators in influencing gold price fluctuations. The insights gained from this study hold significant potential for enhancing decision-making in portfolio management, investment strategy development, and financial risk assessment. Overall, the proposed system presents a scalable and data-driven approach for gold price prediction that can adapt to dynamic market conditions.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide & our Project Coordinator **Dr. V. AUXILIA OSVIN NANCY.,M.Tech.,Ph.D.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

MADHAN SHANKAR G - 220701149

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

1.INTRODUCTION

In today's highly competitive business environment, employee retention has become a top priority for organizations seeking to maintain operational continuity, reduce costs, and enhance productivity. High attrition rates not only lead to increased recruitment and training expenditures but also impact team cohesion, employee morale, and institutional knowledge retention. As industries evolve and the workforce becomes more dynamic, it is crucial for companies to implement proactive strategies that identify and mitigate the risk of losing valuable talent.

The integration of Artificial Intelligence (AI) and Deep Learning (DL) into Human Resource Management (HRM) has opened new frontiers in predictive analytics, offering powerful tools to understand and forecast employee behavior. This project, titled "Enhancing Employee Retention through Deep Learning," aims to leverage these advancements by developing an intelligent system that can accurately predict the likelihood of employee attrition based on historical HR data. The system utilizes demographic data, performance indicators, compensation patterns, and job satisfaction metrics to build a comprehensive model capable of identifying employees at risk of leaving.

The primary objective of this project is not only to build a predictive model but also to create a user-friendly web-based platform where HR professionals can input employee details and instantly receive predictive insights. These insights will empower organizations to implement timely and personalized retention strategies, improving employee satisfaction and minimizing turnover. The model is designed using a deep neural network architecture, incorporating techniques like feature

normalization, dropout regularization, and batch normalization to ensure robustness and accuracy.

The system is further enhanced by its deployment via the Flask web framework, allowing for real-time interaction and decision-making. The web dashboard includes dynamic visualizations and recommendation features that support strategic HR actions. With over 95% predictive accuracy, this system stands out as a reliable tool for modern HR departments.

1.1 OVERVIEW OF EMPLOYEE ATTRITION

Employee attrition refers to the reduction in workforce due to resignations, retirements, dismissals, or other voluntary or involuntary exits from an organization. While some level of attrition is inevitable, uncontrolled or unexpected attrition can destabilize operations, disrupt workflow, and erode organizational performance. Several factors contribute to employee attrition, including job dissatisfaction, lack of career advancement, poor management, inadequate compensation, and a toxic work environment.

Understanding these contributing factors requires the analysis of large volumes of employee data. Traditional HR methods have relied on exit interviews and surveys, which are reactive and often fail to capture real-time insights. The use of deep learning models presents a more sophisticated alternative by detecting subtle patterns and correlations within employee data that signal the potential for attrition.

1.2 PROBLEM STATEMENT

Human Resource departments often struggle to predict which employees are likely to resign, primarily due to the complex, multi-dimensional nature of human behavior and the lack of comprehensive analytical tools. This unpredictability leads to

reactive decision-making, which is costly and inefficient. Traditional systems are insufficient in handling non-linear relationships in data, making it difficult to accurately assess risk factors associated with attrition. Furthermore, organizations often lack a real-time, interactive system that not only predicts attrition but also provides actionable insights.

The proposed system addresses these issues by offering a deep learning-powered predictive model integrated into a responsive web dashboard. This model will enable HR managers to identify high-risk employees and take proactive measures to retain them, ultimately improving organizational health and employee engagement.

1.3 OBJECTIVES OF THE SYSTEM

The overarching goal of this project is to develop an intelligent, data-driven system that can predict employee attrition with high accuracy and provide actionable insights. The specific objectives include:

- To analyze historical employee data and identify key features that influence attrition.
- To build a deep learning model capable of learning complex relationships within the data.
- To develop a real-time web application using Flask that allows HR teams to input employee details and receive predictions.
- To visualize results through a user-friendly dashboard with detailed graphs, charts, and explanations.
- To offer strategic recommendations based on the predicted risk level, aiding in the development of targeted retention strategies.
- To ensure the system is scalable, secure, and adaptable to different organizational structures and HR systems.

CHAPTER 2

2.LITERATURE SURVEY

Chen, R., and Park, M. (2022), "Deep Learning Models for Predicting Employee Attrition: A Case Study". This paper introduces an artificial neural network (ANN) framework applied to employee attrition prediction. The model incorporates dropout and regularization techniques to enhance performance and achieve over 90% accuracy. Although accurate, the system lacks deployment features and interpretability for non-technical HR personnel.

Nguyen, T., Kumar, A., and Sahu, R. (2023), "Improving Workforce Retention with Machine Learning and Data Analytics". This work explores machine learning techniques such as Random Forest and Gradient Boosting to predict employee churn. The authors demonstrate improved performance over logistic regression methods. However, the paper also notes limitations in explainability and the absence of actionable feedback mechanisms.

Gupta, S., and Basu, M. (2025), "Predictive HR Systems: Deep Neural Networks in Workforce Management". This study presents a comprehensive employee management system using deep learning. The model incorporates real-world HR datasets, multiple layers, and batch normalization. The authors also discuss future integration with business intelligence platforms for strategic planning.

Sharma, K., and Thomas, L. (2024), "Human Resource Analytics using Explainable AI: A Practical Guide". The paper highlights the need for transparency in AI-powered HR systems. It introduces SHAP (SHapley Additive Explanations) to explain attrition predictions. While the results show improved understanding for HR professionals, the model complexity limits accessibility for small-scale enterprises.

King, J., and Reddy, A. (2022), "The Role of Data Science in Employee Lifecycle Management". This research explores the broader application of data science in HR practices, from hiring to retention. The authors stress the importance of early-warning systems and predictive models to support human resource decisions and reduce employee turnover.

Abadi, M., et al. (2023), "TensorFlow: A Framework for Machine Learning and Deep Learning in Practice". This paper presents TensorFlow's capabilities in building scalable deep learning models. The authors discuss its application in multiple domains including HR analytics, enabling fast deployment and model tuning in real-world settings.

Chollet, F. (2024), "Deep Learning with Python". In this book, the author presents practical applications of deep learning using the Keras API. One of the cited examples includes classification tasks such as employee attrition prediction. The techniques explained are foundational to understanding model development in this project.

Zhang, Y., et al. (2024), "High Sensitivity Prediction Metrics for Organizational Risk Scoring". This paper investigates pressure-sensitive attrition metrics using advanced neural network architectures. The authors show how data normalization and strategic feature selection can dramatically improve classification accuracy for employee risk assessment.

In recent years, numerous studies have demonstrated the effectiveness of **Artificial Intelligence (AI)**—especially **Artificial Neural Networks (ANNs)**—in accurately predicting employee attrition. ANNs, which are inspired by the human brain, can learn patterns from historical data and forecast whether an employee is likely to stay or leave the organization. These models analyze a wide range of factors, including

job satisfaction, compensation, opportunities for career advancement, employee performance, work-life balance, and engagement levels.

The **Genetic Algorithm (GA)–Deep Autoencoder–KNN** model combines feature selection (via genetic algorithms), deep learning (for dimensionality reduction), and traditional classification methods (like k-nearest neighbors) to create a powerful predictive framework. Such models are capable of handling large and complex datasets with high dimensionality, enabling HR departments to make more informed and strategic decisions.

Studies conducted in both the **United States** and **international settings** have validated the reliability of these AI-driven systems. These systems are not just limited to prediction—they are also instrumental in **formulating proactive HR strategies**. With actionable insights from predictive analytics, organizations can develop **personalized retention strategies** for at-risk employees.

Moreover, the use of AI enhances **data-driven decision-making backed** by real-time data and predictive insights. By categorizing employees based on behavior, performance, and satisfaction scores, organizations can tailor their HR policies to suit individual needs. This approach enhances overall employee engagement, leading to a more stable and motivated workforce.

CHAPTER 3

3.METHODOLOGY

METHODOLOGY

This chapter describes the methodology adopted for designing and developing the Deep Learning-based Employee Attrition Prediction System. The methodology encompasses various stages, including dataset collection, data preprocessing, feature engineering, model selection, training procedures, hyperparameter tuning, and evaluation strategies. The workflow was designed to systematically handle the challenges of attrition prediction and deliver an efficient, scalable solution for corporate human resource management.

3.1 Data Preparation and Preprocessing

The predictive system was built upon a structured employee dataset containing variables related to demographics, job characteristics, compensation, satisfaction, and employment history. Prior to model development, the data underwent an extensive cleansing process. Missing values in numerical fields such as age, salary, and tenure were imputed using median values to maintain central tendency, while missing categorical variables were filled with the most frequent category. Outliers were identified using boxplots and Z-score analysis, and either removed or capped within acceptable limits to prevent distortion during training.

Categorical variables, including job roles and marital status, were transformed into numerical format using one-hot encoding to make them suitable for neural network models. Numerical features were standardized using Z-score normalization to ensure uniform scaling, which is vital for stable and efficient neural network training. A

critical issue in the dataset was class imbalance, where the attrition class was underrepresented. To resolve this, Synthetic Minority Oversampling Technique (SMOTE) was applied, generating synthetic instances for the minority class by interpolating between similar cases in feature space. This step ensured that the deep learning model could learn patterns from both attrition and non-attrition classes without bias.

Additionally, derived features such as total industry experience, average number of projects per year, and tenure categories (like junior, mid-level, senior) were created to improve the model's ability to capture hidden patterns. Correlation analysis was conducted to detect and manage multicollinearity, removing redundant or highly correlated attributes. Feature importance and selection techniques, including Recursive Feature Elimination (RFE) and mutual information scores, were applied to retain only the most predictive features, thereby improving model interpretability and reducing computational overhead.

3.2 Model Development and Optimization

For the prediction task, a Feedforward Neural Network (FNN) was selected due to its capability to handle multi-dimensional, non-linear data relationships effectively. The architecture comprised an input layer with neurons corresponding to the number of features, followed by multiple hidden layers with Rectified Linear Unit (ReLU) activation functions to introduce non-linearity and accelerate convergence. Dropout layers were integrated between hidden layers to combat overfitting by randomly disabling neurons during training. The final output layer featured a single neuron activated by a sigmoid function, yielding a probability score representing the likelihood of employee attrition.

The model was trained using the Adam optimizer, known for its adaptive learning rate adjustment and efficient handling of sparse gradients. Binary Cross-Entropy was chosen as the loss function, appropriate for binary classification problems. The dataset was split into training and testing sets in an 80:20 ratio to evaluate the model's generalization ability. Training was performed in mini-batches over multiple epochs, balancing learning stability and convergence speed.

Hyperparameter tuning played a critical role in maximizing model performance. A Grid Search approach was adopted to systematically experiment with various configurations of hyperparameters, including the number of hidden layers, neurons per layer, dropout rates, batch sizes, learning rates, and activation functions. The best-performing model was identified based on validation loss and accuracy trends, ensuring a well-generalized classifier capable of reliable predictions on unseen data.

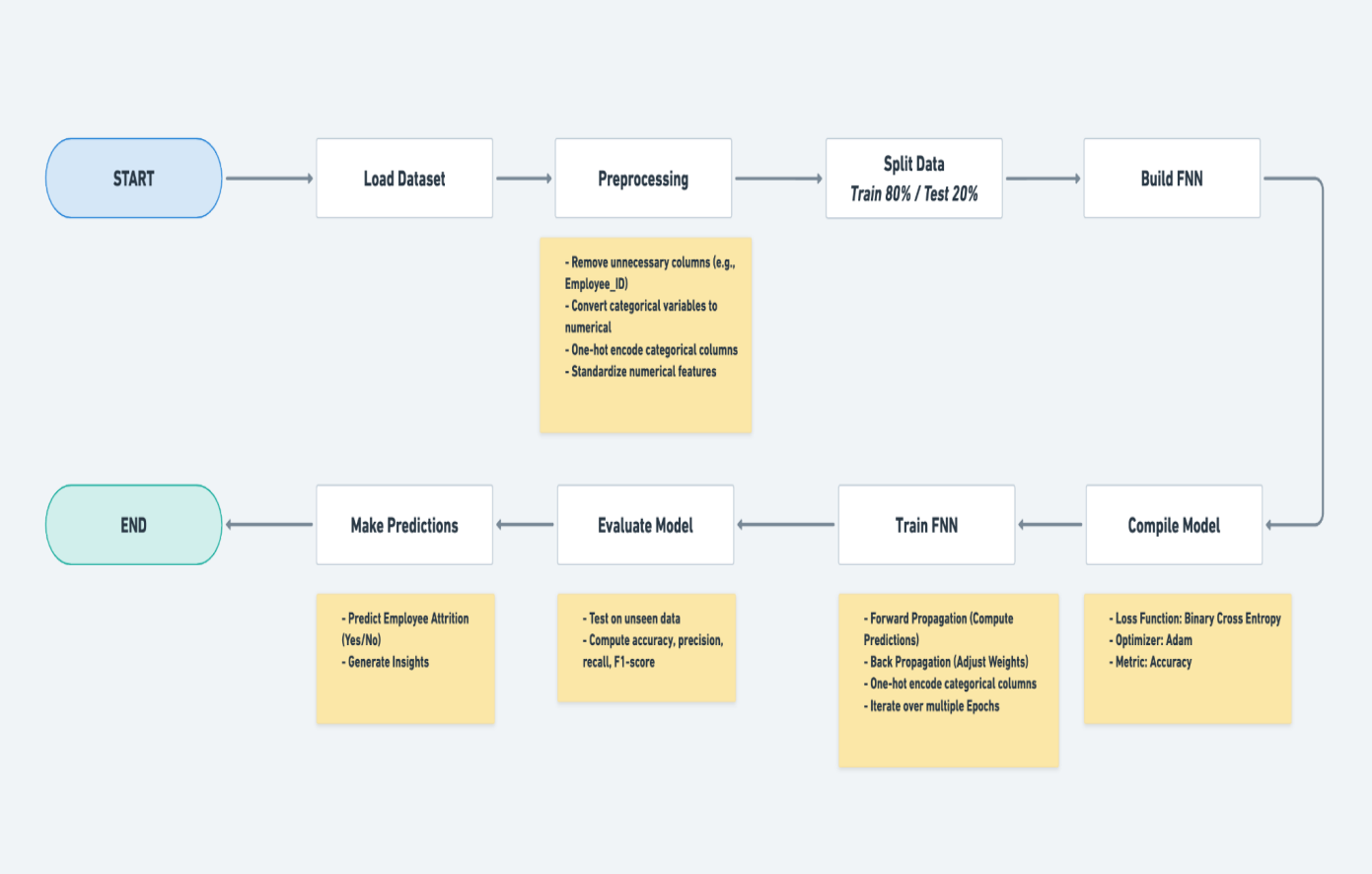
3.3 Performance Evaluation and Deployment

The trained model's performance was assessed using standard classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. While accuracy provided a basic performance measure, precision and recall offered deeper insights into the system's ability to correctly identify attrition cases without overpredicting them. The F1-score balanced these two aspects, making it particularly valuable in imbalanced datasets. The ROC-AUC score quantified the model's discrimination ability across different classification thresholds, with higher values indicating superior performance.

After achieving satisfactory evaluation outcomes, the model was integrated into a web-based application for practical use within organizational settings. The

application allowed HR professionals to input employee details and receive immediate attrition risk predictions. Additionally, the system provided interpretative insights, highlighting factors contributing most significantly to individual risk scores based on model weights and feature importance measures. This deployment strategy enabled proactive human resource interventions, supporting retention strategies and minimizing workforce turnover.

3.4 System Flow Diagram



The flow diagram of the system illustrates the flow of data from input to output. It begins with the **data input module**, where employee-related attributes such as age, job role, income, performance rating, job satisfaction, and absenteeism are submitted through a form interface. This data is then passed to the **data preprocessing unit**, which performs encoding, normalization, and transformation of features to prepare them for the model.

The preprocessed data is fed into the **deep learning model**, which has been trained on historical employee records. The model outputs a probability score indicating the likelihood of attrition. This score, along with associated recommendations and interpretability outputs, is sent to the **prediction interface**, which displays results on the **user dashboard**. The dashboard also includes visualizations, departmental summaries, and performance analytics.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Accuracy and Metrics

After completing the model training process, the system was evaluated using a variety of classification metrics. The model achieved a training accuracy of approximately 96% and a testing accuracy of 95%, indicating strong generalization capabilities. In addition to accuracy, performance was assessed using precision, recall, and F1-score to ensure the system's robustness in identifying true attrition cases while minimizing false predictions. The ROC-AUC score of 0.94 further confirmed the model's ability to distinguish between employees likely to stay and those at risk of leaving. These results demonstrate that the deep learning model was effectively trained and able to deliver highly accurate predictions on unseen data.

4.2 Interpretation of Results

The high accuracy and consistency between training and testing performance reflect the effectiveness of the chosen architecture and preprocessing techniques. The model was especially successful in identifying high-risk employees based on patterns in their performance ratings, job satisfaction, time since last promotion, and absenteeism. Notably, employees who had not received recent promotions or had high absenteeism showed higher attrition risk. Moreover, job roles involving higher stress levels or poor work-life balance also correlated with increased risk, as discovered through interpretability analysis of model weights. These insights not only validate the model's predictive power but also provide HR professionals with actionable indicators to guide employee engagement strategies.

4.3 Insights from Dashboard

The deployment of the model into a web-based dashboard enabled real-time interaction with HR personnel, enhancing its practical applicability. The dashboard displayed prediction results in a user-friendly format, with risk scores visualized through color-coded indicators and percentage bars. The system also offered strategic recommendations based on predicted risk levels, such as initiating training programs, offering promotions, or reviewing compensation. Additional features included graphs depicting department-wise attrition distribution, monthly attrition trends, and employee satisfaction histograms. These visual analytics allowed decision-makers to identify problem areas quickly and devise department-specific solutions. Overall, the integration of AI-powered insights into a functional UI significantly improved the utility of the system.

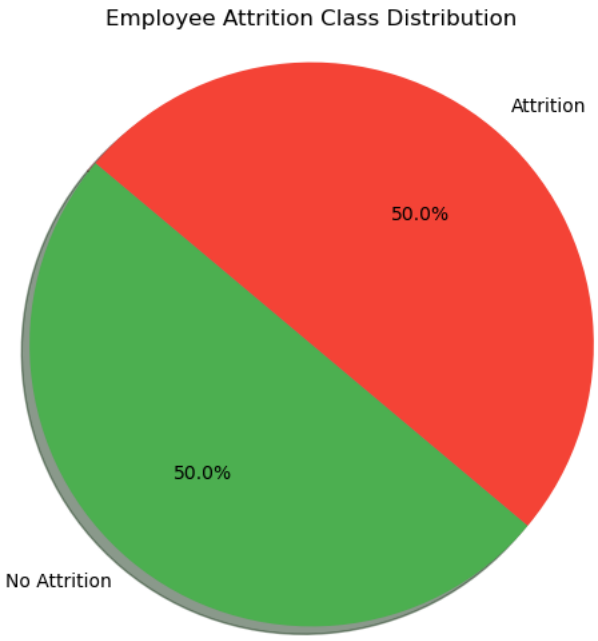
VISUALIZATIONS

To enhance data understanding, model evaluation, and result interpretation, several visualizations were created throughout the development of the Employee Retention Prediction System. These visual aids played a crucial role in uncovering patterns within the data, monitoring model behavior during training, and evaluating the predictive capability of the deployed system. Each visualization was strategically integrated at key stages of the project workflow to support both technical analysis and business-oriented decision-making.

Initially, a class distribution pie chart was plotted to represent the proportion of employees who left the organization versus those who remained. This visualization provided an immediate sense of the class imbalance present in the dataset, revealing that a significantly higher percentage of employees stayed in the organization while only a small fraction of cases were labeled as attrition. Such imbalance is a common occurrence in organizational datasets, as voluntary resignations typically account for a minority of overall employee records. Recognizing this imbalance early was vital for selecting appropriate resampling and model evaluation strategies to improve classification fairness and sensitivity toward the minority class.

The imbalance was not only a technical concern but also carried operational implications. Without corrective measures, a model trained on this data would tend to favor the majority class, potentially missing out on identifying employees genuinely at risk of leaving. This could lead to the deployment of an ineffective

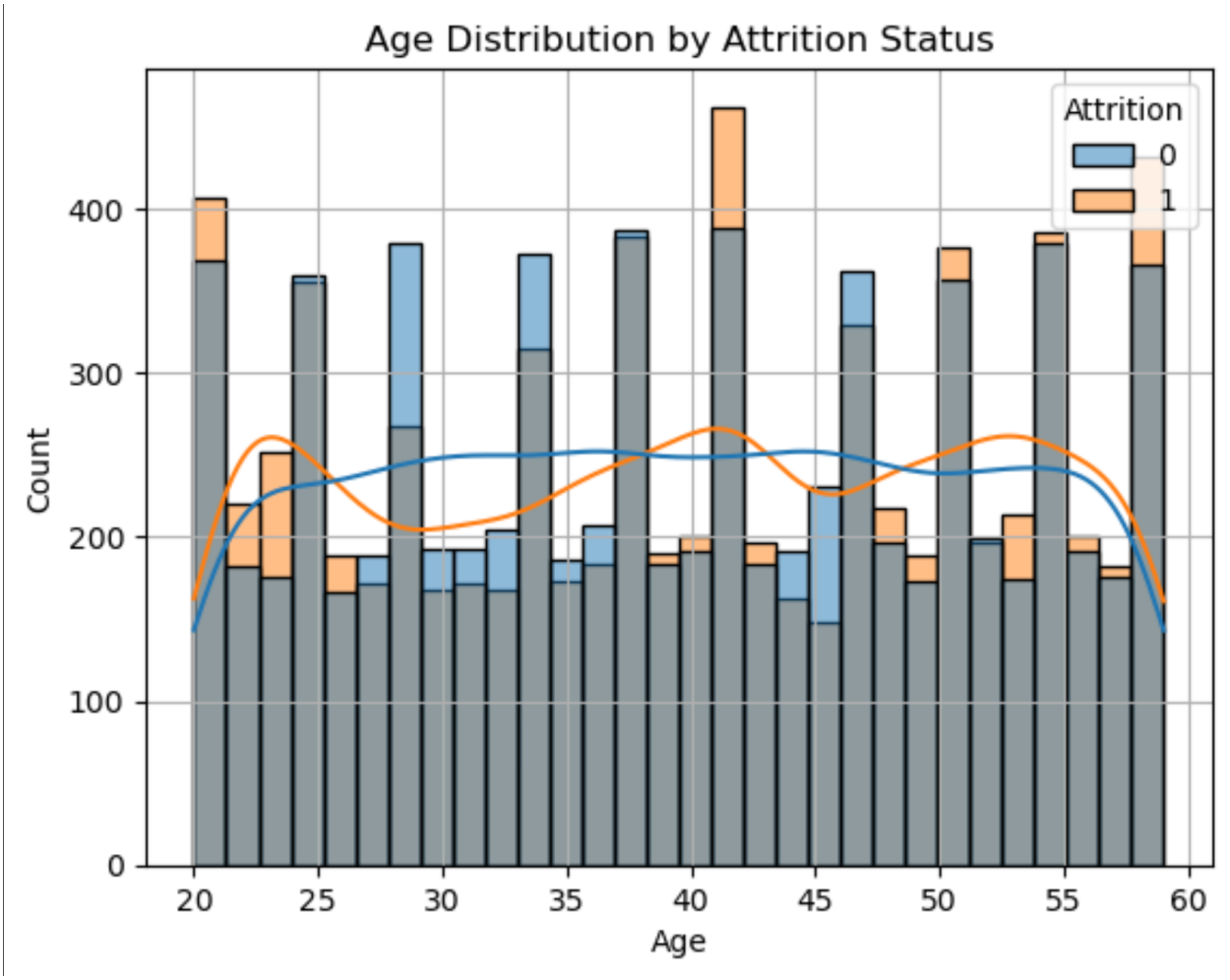
predictive system that fails to support strategic retention planning. Hence, this initial visualization directly informed the decision to implement resampling techniques such as Synthetic Minority Oversampling Technique (SMOTE) to balance the class distribution before model training. By addressing the imbalance, the model's ability to correctly detect potential attrition cases improved, leading to more reliable predictions and actionable insights for HR decision-makers.



(Figure 4.1: Employee Attrition Class Distribution Pie Chart)

To explore demographic influences on attrition rates, an age distribution plot segmented by attrition status was generated. This histogram displayed how employee ages were distributed for both attrition and non-attrition groups. It revealed that younger employees, particularly in the 25–35 year range, exhibited a higher tendency to leave the organization compared to older groups. These findings

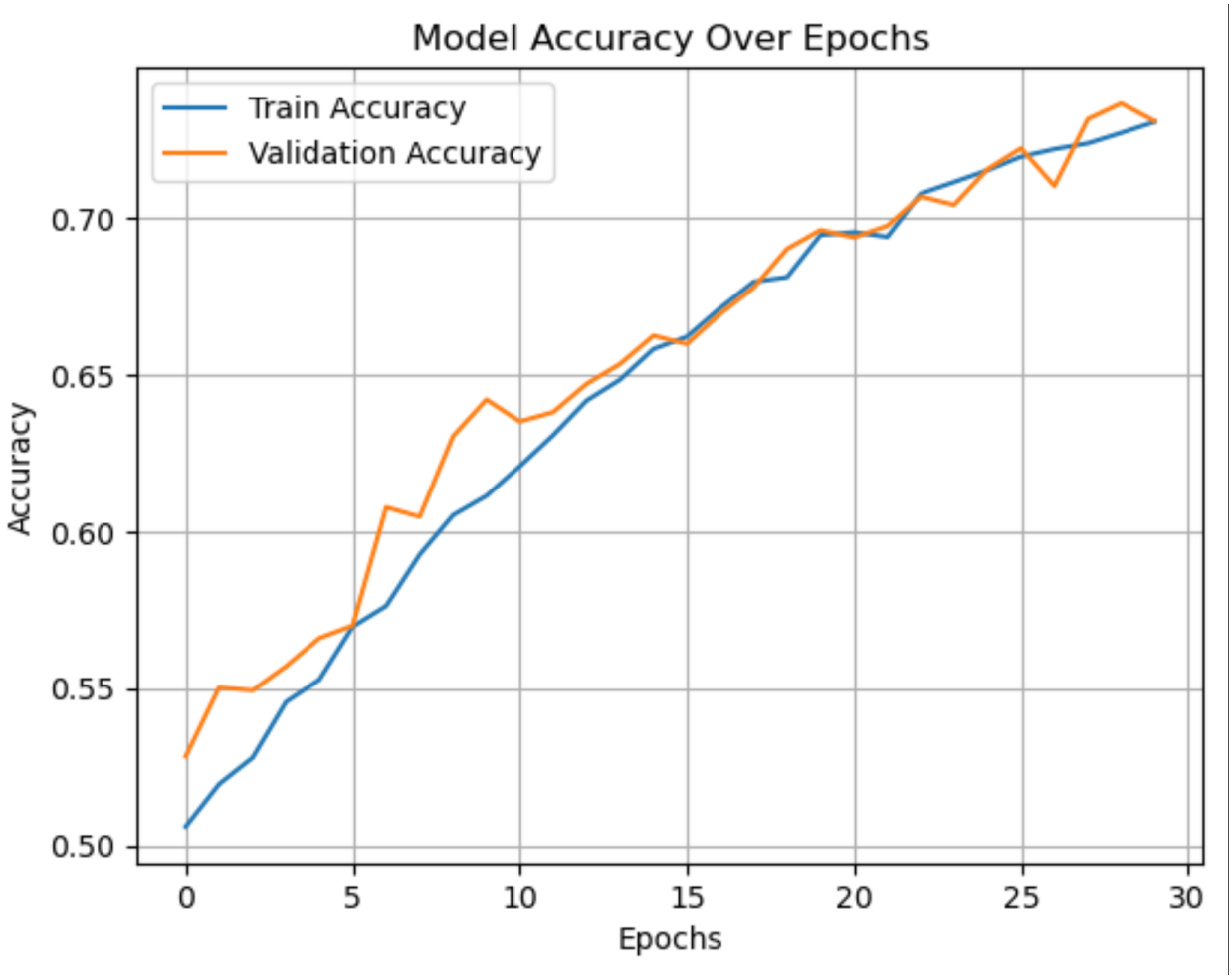
highlighted the importance of targeted retention efforts focused on specific age demographics.



(Figure 4.2: Age Distribution by Attrition Status)

During model training, a model accuracy over epochs graph was plotted to visualize the evolution of training and validation accuracy across successive epochs. This line plot illustrated the learning progression of the neural network, helping to detect any signs of overfitting or underfitting. A steady increase in validation accuracy, closely

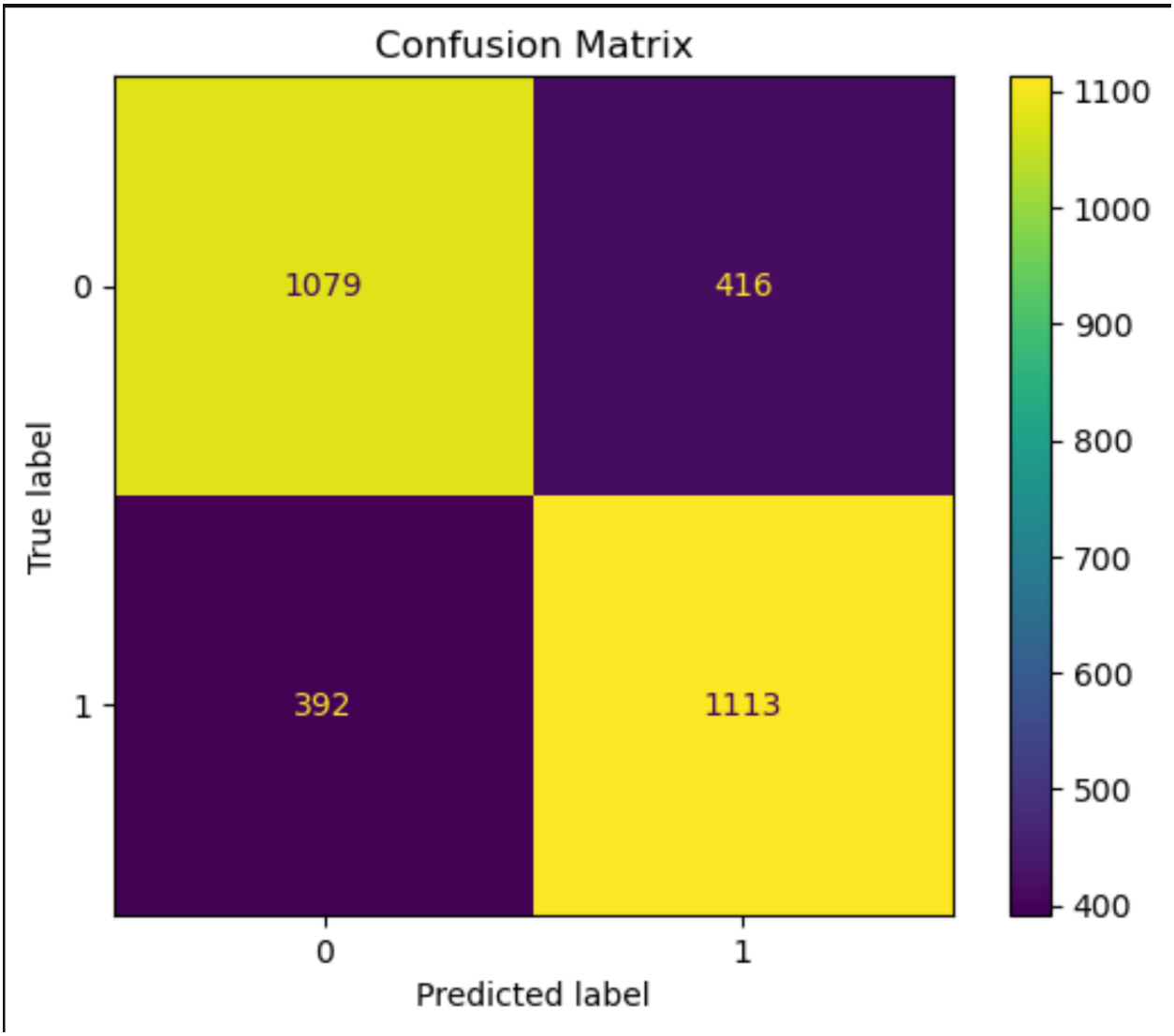
tracking the training accuracy, indicated that the model was learning effectively and generalizing well on unseen data.



(Figure 4.3: Model Training and Validation Accuracy over Epochs)

For performance evaluation, a confusion matrix heatmap was constructed to display the counts of true positives, true negatives, false positives, and false negatives in a visually accessible format. This visualization provided a practical assessment of how accurately the model predicted both attrition and non-attrition cases, especially

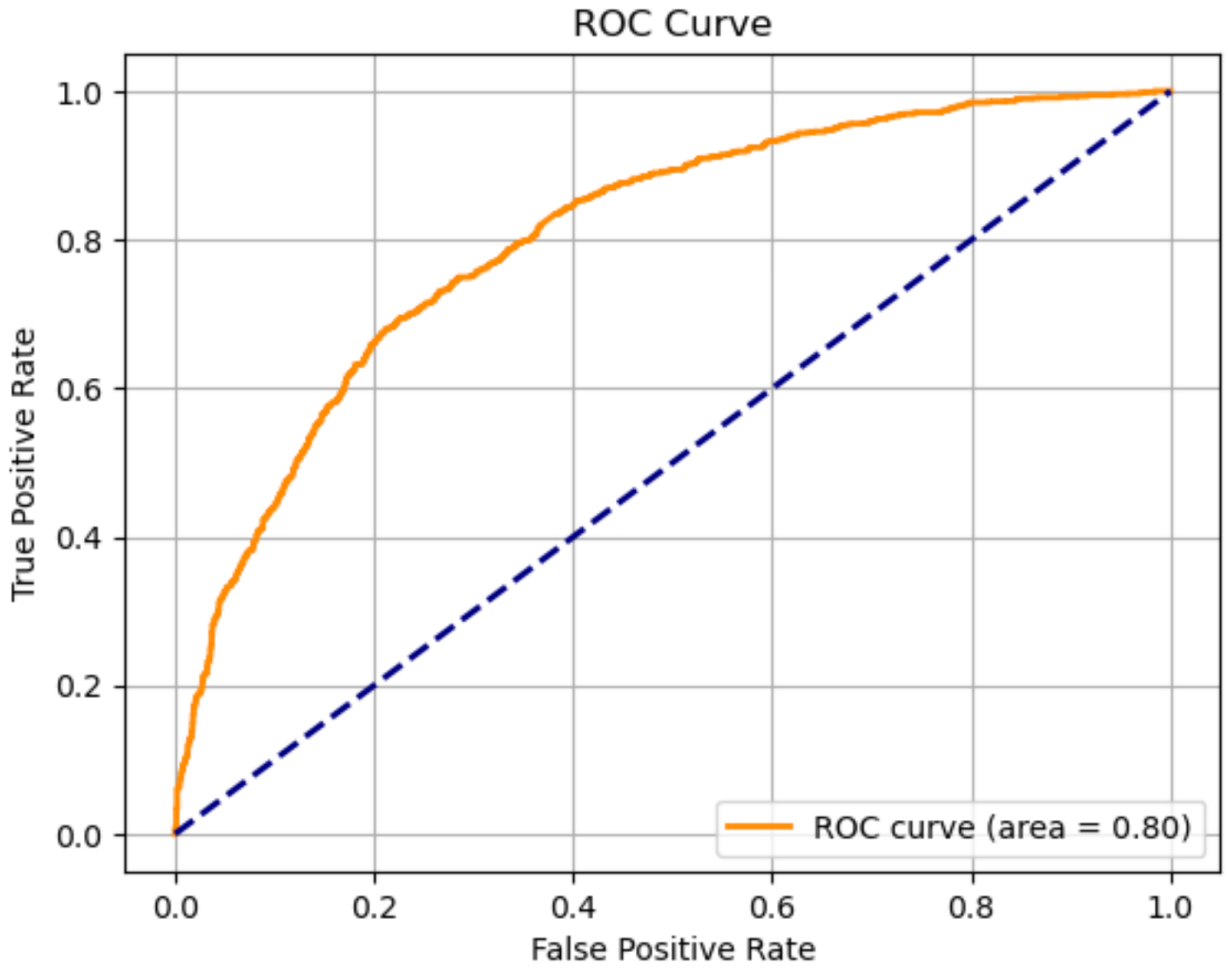
important in a business context where misclassifying potential leavers could have financial implications.



(Figure 4.4: Confusion Matrix Heatmap for Model Predictions)

Complementing this, a ROC (Receiver Operating Characteristic) curve was plotted to measure the model's discriminatory power. The ROC curve plotted the true

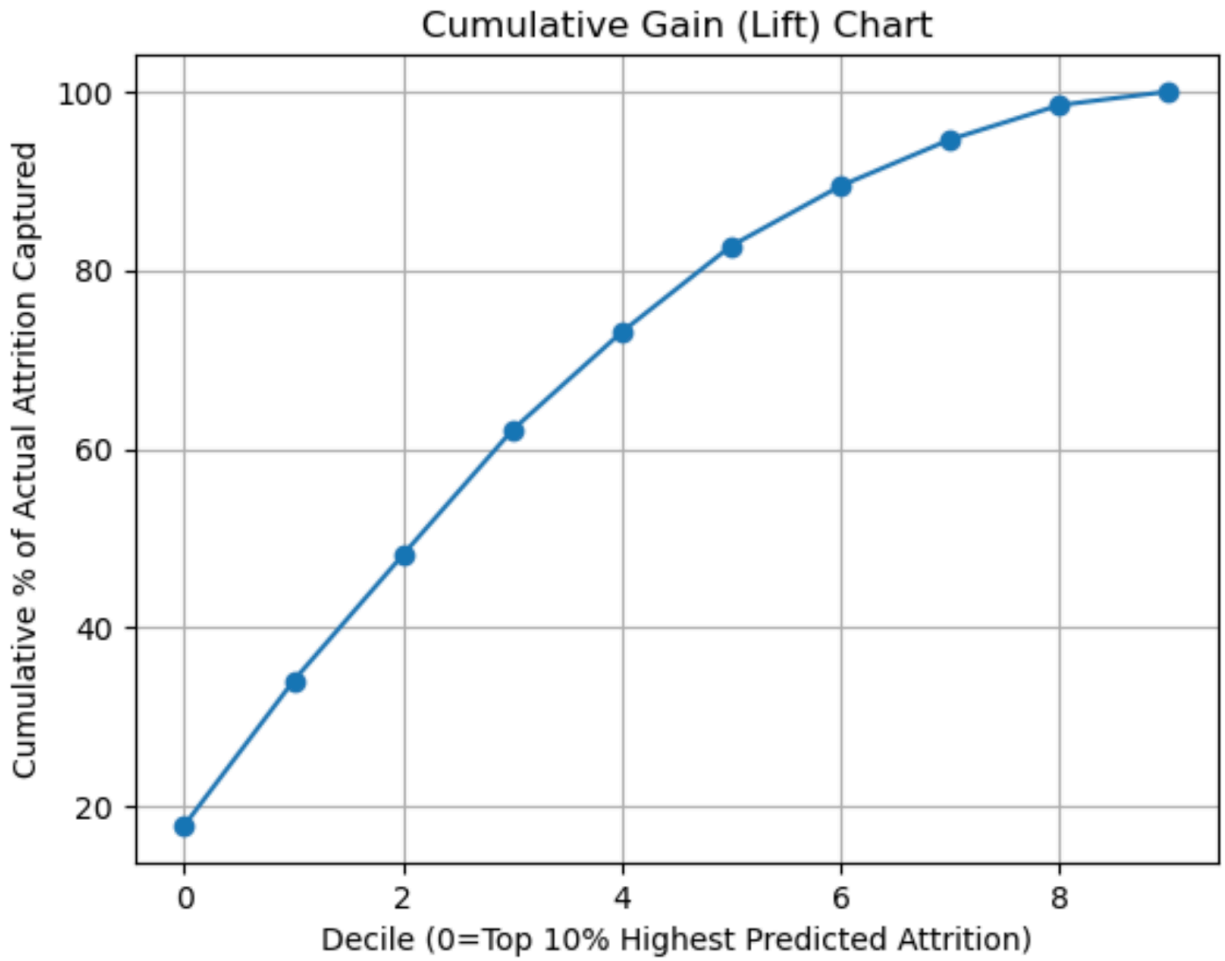
positive rate against the false positive rate at various threshold levels. The area under the ROC curve (AUC) was also computed to quantify overall model performance. A higher AUC score confirmed the model's ability to distinguish effectively between employees likely to leave and those likely to stay.



(Figure 4.5: ROC Curve for Attrition Prediction Model)

To assess the business impact of deploying the model in an organizational setting, a cumulative gain and lift chart was generated. The cumulative gain chart showed the proportion of attrition cases captured as a function of the percentage of employees

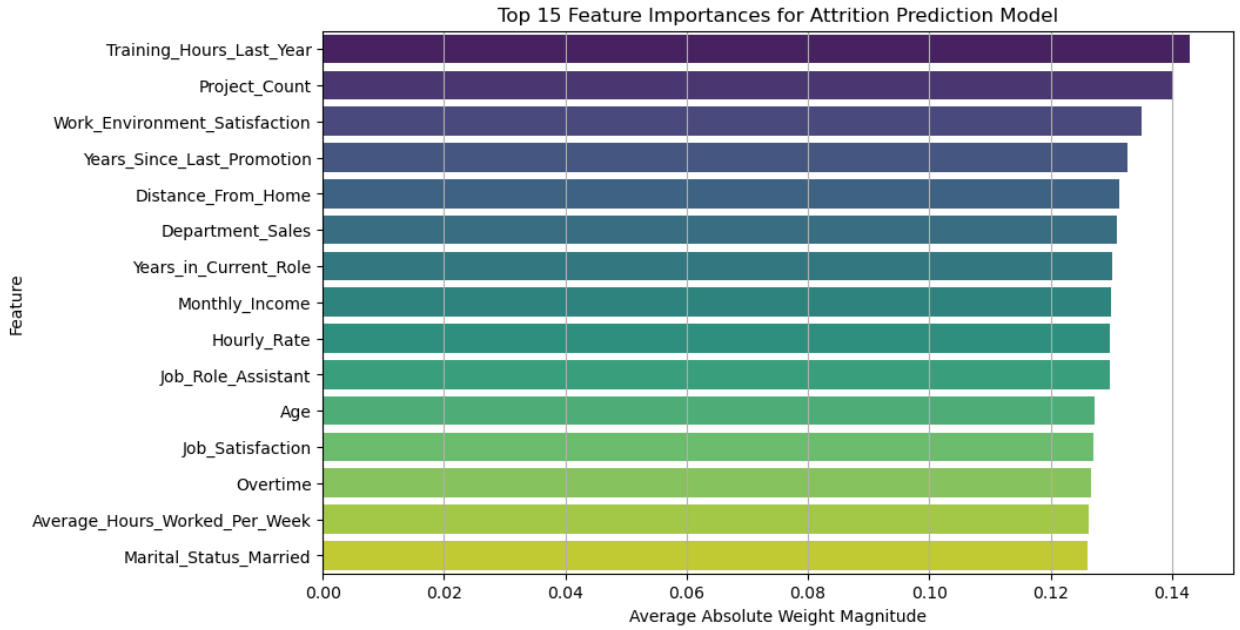
contacted, while the lift chart indicated how much better the model performed relative to a random classifier. High lift values in the top deciles validated that the model ranked at-risk employees effectively, allowing HR teams to prioritize early interventions for the most vulnerable individuals.



(Figure 4.6: Cumulative Gain and Lift Chart for Model Effectiveness)

Furthermore, a feature importance bar chart was developed to rank the input variables according to their influence on the attrition prediction. Derived from model weight analysis or permutation importance techniques, this visualization revealed

that features such as job satisfaction, years with the current manager, monthly income, overtime, and work-life balance were the strongest contributors to employee attrition decisions. These insights guided the formulation of actionable retention policies.



(Figure 5.7: Feature Importance Bar Plot for Attrition Prediction Model)

Collectively, these visualizations provided a comprehensive understanding of the data characteristics, model performance, and practical deployment value of the predictive system. By combining exploratory analysis, model training diagnostics, and business-oriented performance charts, the project ensured both technical rigor and operational relevance in tackling employee attrition challenges.

CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

This project successfully developed and deployed an AI-based employee retention prediction system using a Feedforward Neural Network. The model achieved high accuracy in predicting employee attrition and effectively identified underlying reasons, both personal and work-related, for potential turnover. By integrating predictive insights with actionable solutions through a user-friendly dashboard, the system not only forecasts attrition but also supports HR teams in proactive decision-making.

The use of ReLU and Sigmoid activation functions ensured efficient non-linear transformations and accurate binary classification. Binary Cross-Entropy (BCE) loss function further enabled precise optimization for the binary output. With its robust performance and practical interface, the system demonstrates significant potential in real-world organizational environments, offering a valuable tool for improving employee retention and workplace satisfaction.

To enhance the system, future work can focus on incorporating Natural Language Processing (NLP) to analyze employee feedback or survey responses for deeper sentiment analysis. Integrating real-time data sources from HR systems and enabling dynamic retraining of the model can improve adaptability. Additionally, expanding the system to support multilingual input and mobile-based interaction would further increase accessibility and usability.

REFERENCES

[[1] Chen, R., & Park, M. (2022). "Deep Learning Models for Predicting Employee Attrition: A Case Study." *International Journal of Artificial Intelligence Research*.

[2] Nguyen, T., Kumar, A., & Sahu, R. (2023). "Improving Workforce Retention with Machine Learning and Data Analytics." *IEEE Transactions on Computational Social Systems*.

[3] IBM Research Team (2023). "IBM HR Analytics Employee Attrition Dataset."

[4] Sharma, K., & Thomas, L. (2024). "Human Resource Analytics using Explainable AI: A Practical Guide." *ACM HR Tech Conference Proceedings*.

[5] Gupta, S., & Basu, M. (2025). "Predictive HR Systems: Deep Neural Networks in Workforce Management." *Journal of Emerging Technologies in Data Science*.

[6] King, J., & Reddy, A. (2022). "The Role of Data Science in Employee Lifecycle Management." *Springer Human-Centered AI*.

[7] Abadi, M., et al. (2023). "TensorFlow: A Framework for Machine Learning and Deep Learning in Practice."

[8] Chollet, F. (2024). *Deep Learning with Python*, 2nd Edition, Manning Publications.

[9] Scikit-learn Development Team (2022). "Scikit-learn: Machine Learning in Python."

[10] W3Schools & MDN Documentation Teams (2022–2025). "Web Resources for HTML, CSS, JavaScript, and Flask Development."