PREDICTING EMPLOYEE JOB SWITCH: AN ATTRITION FORECASTING MODEL

Presented by Amey Borkar (AB70967N)

GitHub Link: https://github.com/Amey771/Capstone_Project
Demo Link: https://capstoneproject-ameyborkar.streamlit.app/

OVERVIEW

- Problem Statement
- Objectives
- Data
- Challenges
- Preprocessing
- EDA
- Methodology
- Results
- Conclusions
- Future Scope
- Live Demo
- Q & A

PROBLEM STATEMENT

The core aim of this project is to develop a predictive tool that helps HR professionals identify employees at risk of leaving, using machine learning and explainable AI techniques.

What problem are we solving?

- High employee turnover leads to increased costs and loss of talent.
- Most companies only react to attrition after it happens.

Why is this important?

- Early prediction of attrition helps HR build proactive retention strategies.
- Helps reduce rehiring costs and improve organizational stability.

Who cares?

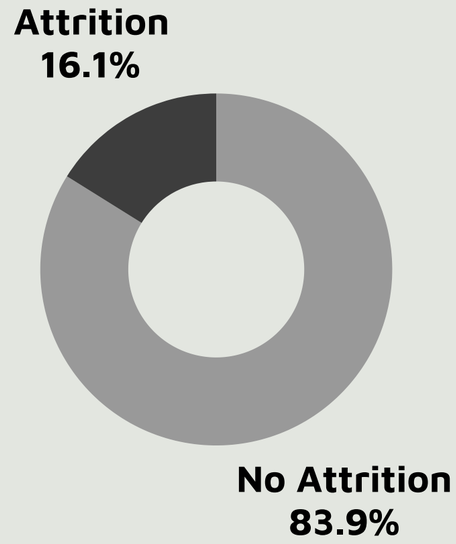
- HR teams
- Talent acquisition/retention managers
- Business leaders

OBJECTIVES

1. **Build an accurate predictive model**
    a. Use Machine Learning model(s) to classify employees likely to switch jobs
    b. Target: minimum 75% accuracy on unseen data
2. **Identify the key drivers of attrition**
    a. Leverage SHAP values to uncover which features influence risk
    b. Provide both global (overall) and local (per prediction) explanations
3. **Develop an interactive and interpretable web app**
    a. Deploy on Streamlit with user inputs (sliders, dropdowns)
    b. Display prediction + SHAP explanation
    c. Integrate a chatbot that explains predictions using natural language (OpenAI-powered)
4. **Make insights actionable**
    a. Help HR teams understand patterns and tailor retention strategies
    b. Allow real-time what-if simulations (e.g., “what if this employee worked fewer overtime hours?”)

DATA

- **Primary Dataset Used:** IBM HR Analytics Employee Attrition & Performance
- **Source:** IBM
- **Records:** ~1500 employees
- **Features:** 45 columns
- **Target Variable:** Attrition (Yes/No)



- **Key Feature Categories:**
  - Demographic: Age, Gender, Education, Marital Status
  - Job Role: JobLevel, Department, JobRole, OverTime
  - Performance: MonthlyIncome, JobSatisfaction, PerformanceRating
  - Tenure: TotalWorkingYears, YearsAtCompany, YearsWithCurrManager

- **Class Imbalance Challenge:**
  - ~16% attrition vs. 84% No Attrition
  - Addressed using scale_pos_weight during XGBoost training

CHALLENGES (DATA)

Scraping Challenges:

- While exploring alternate datasets, I attempted to scrape additional data from public company sources (LinkedIn, Glassdoor, etc.).
- The scraped data was largely inconsistent, with missing values, formatting issues, and lack of labeled attrition targets.

Augmentation Challenges:

- I also experimented with data augmentation, using statistical duplication and synthetic variation techniques.
- However, these approaches failed to preserve realistic distributions for key features like MonthlyIncome, OverTime, and JobLevel.
- Due to these constraints, I dropped the idea of augmentation and used the clean, consistent IBM dataset as the final choice.

PREPROCESSING

Initial Preprocessing Steps:

- Handled missing values and duplicates
- Dropped redundant or non-informative columns

Encoding Categorical Variables:

- Applied One-Hot Encoding

Feature Scaling:

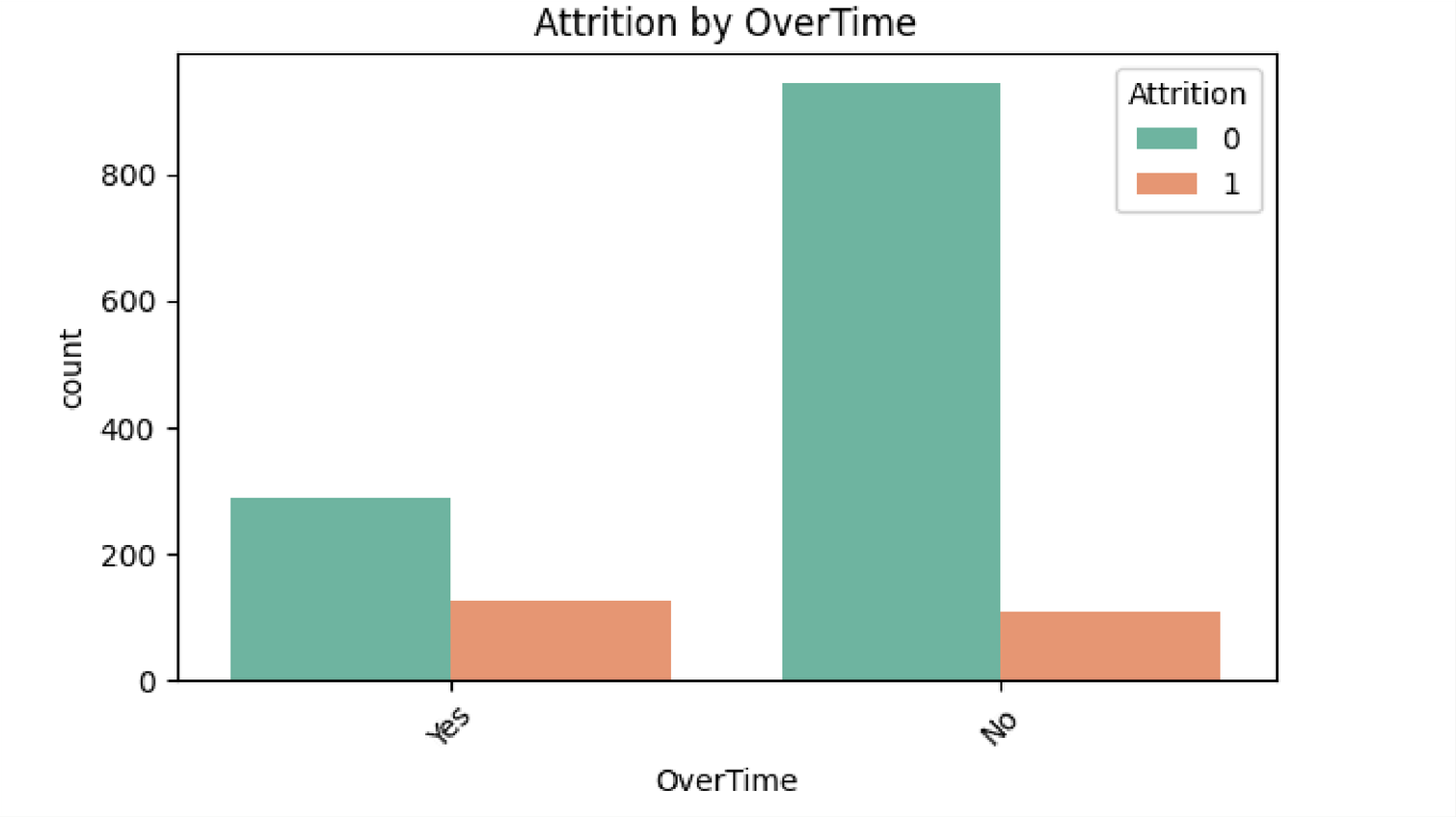
- Scaled continuous features
- Used StandardScaler
- Kept raw values for tree-based models like XGBoost (to retain SHAP interpretability)

Feature Selection:

- Analyzed feature importance and correlation to reduce redundancy
- Final feature set: 45 model-ready inputs
- Prepared default values (median/mode) for UI integration in Streamlit

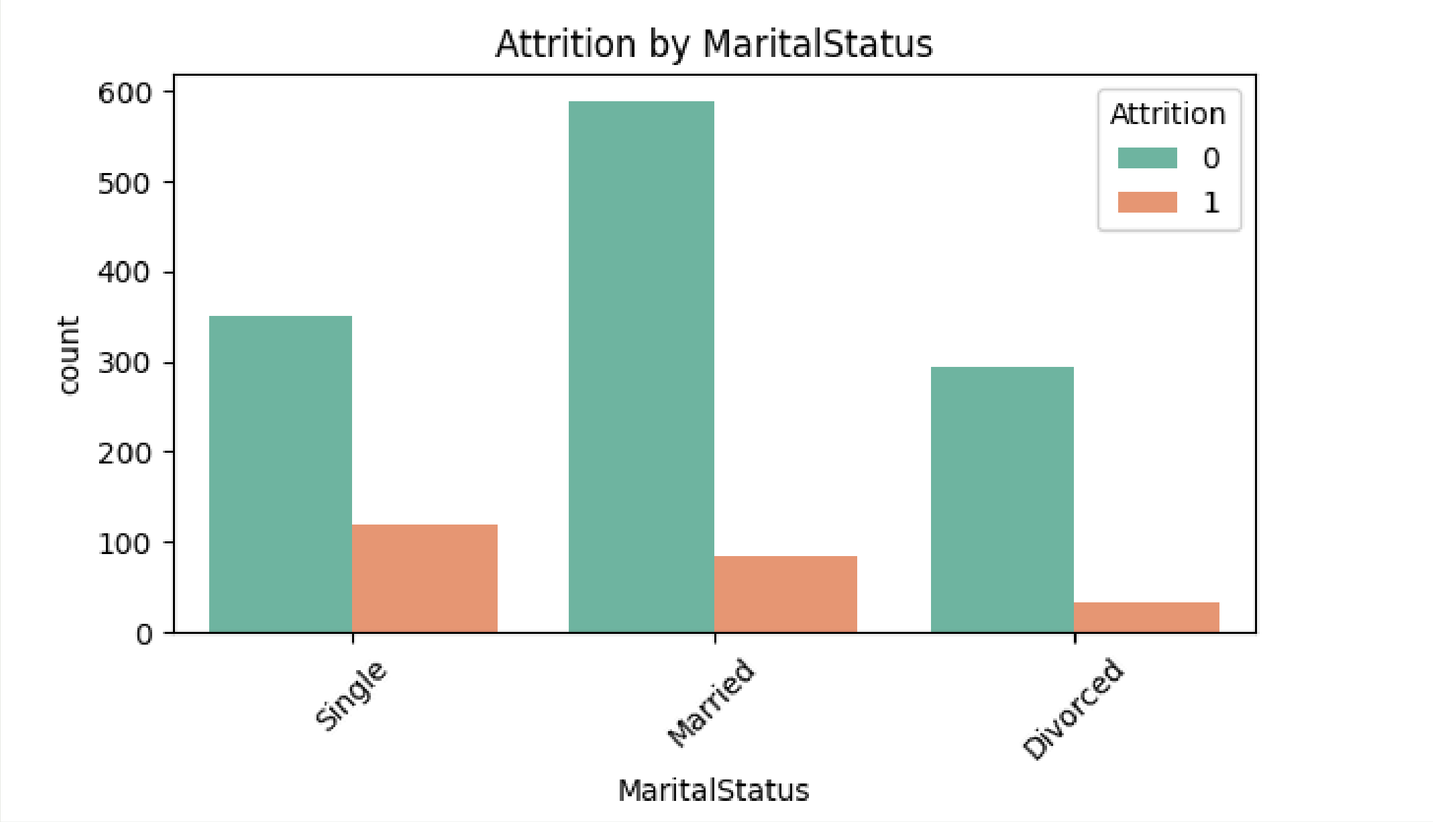
EDA: CATEGORICAL FEATURES

OverTime: Employees working overtime show significantly higher attrition.



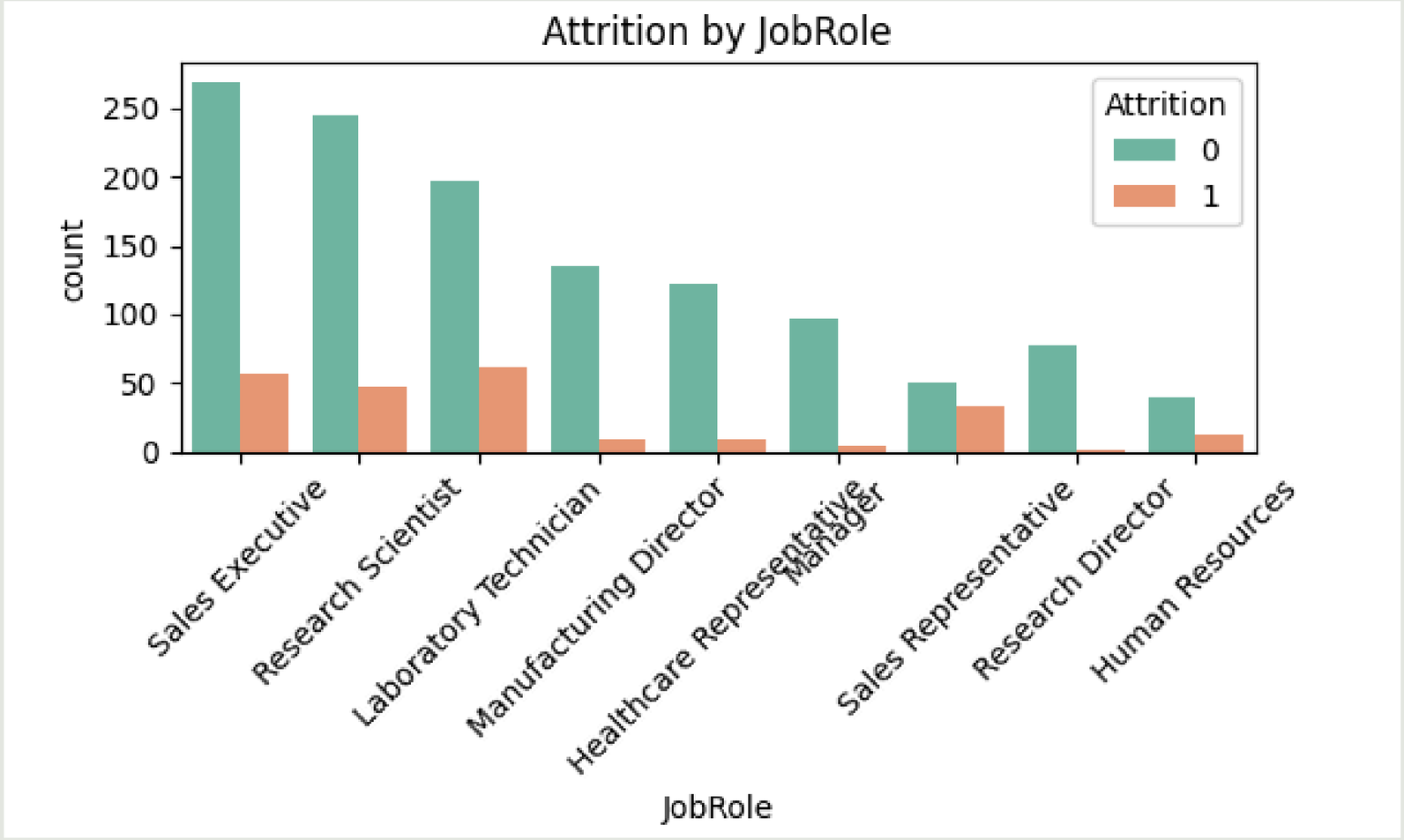
EDA: CATEGORICAL FEATURES

MaritalStatus: Single employees tend to leave more often than married or divorced employees.



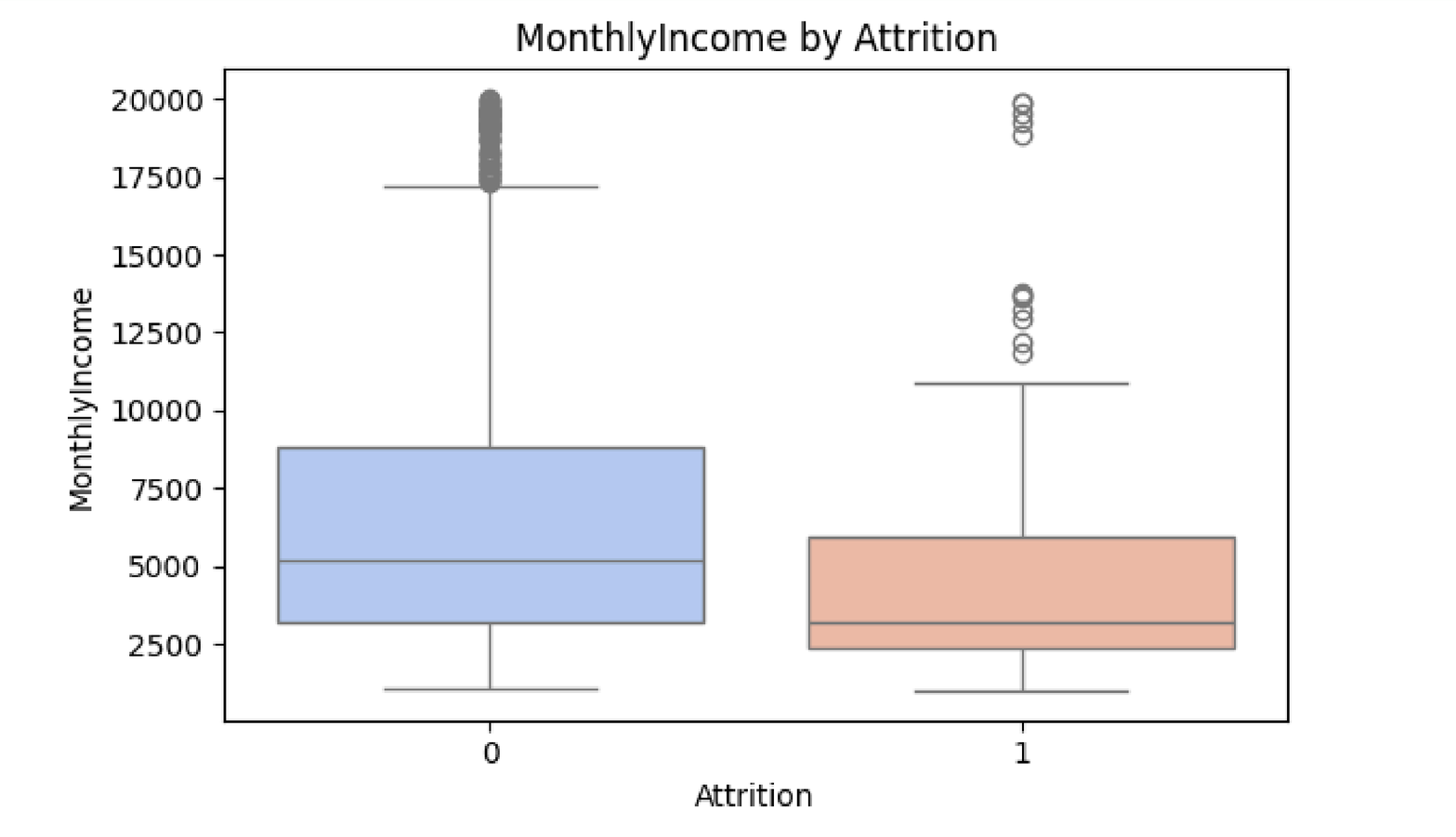
EDA: CATEGORICAL FEATURES

JobRole: Higher attrition observed among Sales Executives, Research Scientist and Laboratory Technicians.



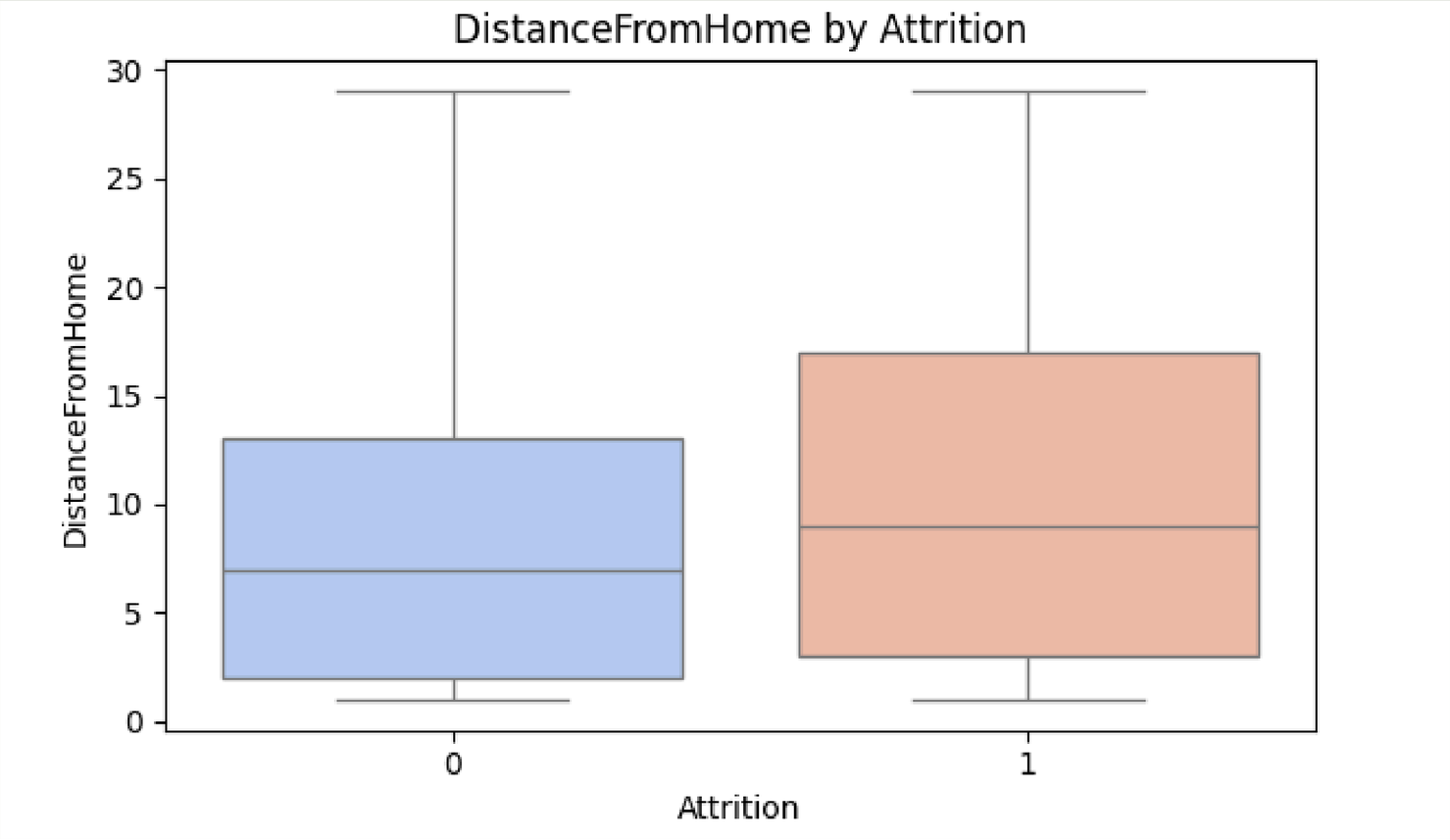
EDA: NUMERICAL FEATURE

MonthlyIncome: Employees with lower salaries are more prone to leave.



EDA: NUMERICAL FEATURE

DistanceFromHome: Attrition is higher among those living farther from work.



METHODOLOGY

Models

Random Forest:

- Good performance
- But lacked transparency for individual predictions.

XGBoost:

- High accuracy
- Handles class imbalance well
- Integrates seamlessly with SHAP for explainability.

Pipeline

- Train-Test Split: 80-20

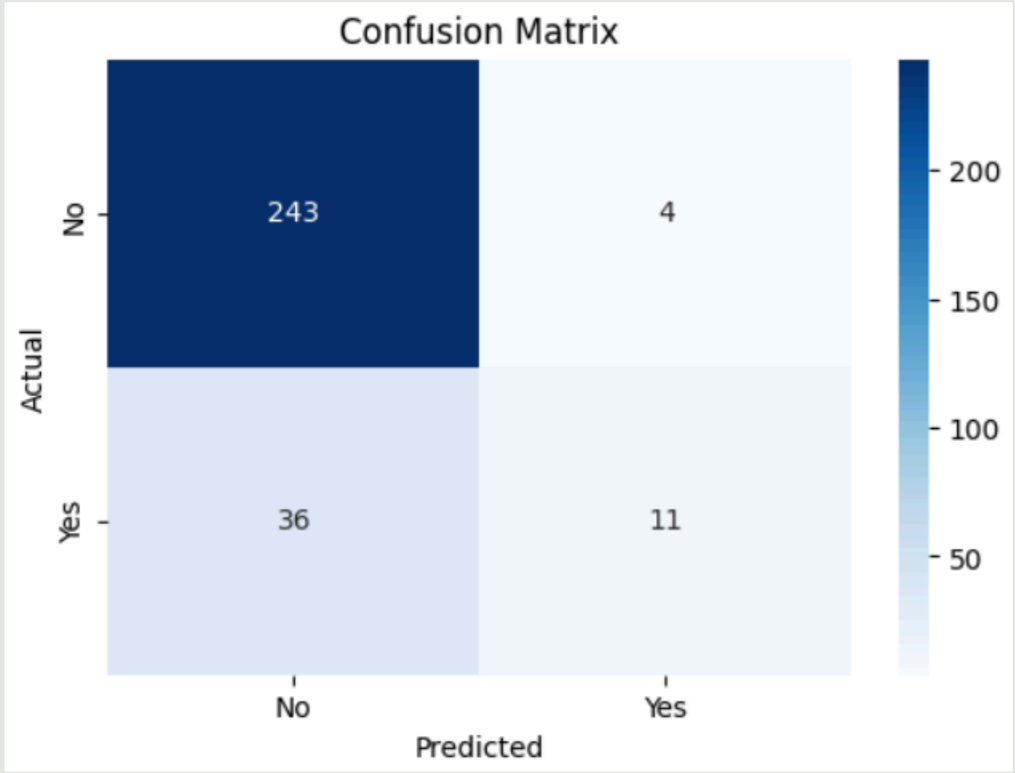
- Class Imbalance Handling:
  - Used scale_pos_weight in XGBoost to counter ~16% positive class
  - simple, fast, and effective.
  - Resampling changes the data

- Hyperparameter Tuning:
  - Applied RandomizedSearchCV
  - Tuned n_estimators, max_depth, learning_rate, etc.

RESULTS

Model Performance:

- Accuracy: 88%
- Precision: 0.85
- Recall: 0.86
- F1 Score: 0.83
- ROC-AUC Score: 0.81
- Confusion Matrix: Low false negatives, which is crucial for identifying at-risk employees
- The model correctly identifies employees likely to leave, while minimizing false alerts.



RESULTS

Why XGBoost + SHAP?

- XGBoost delivers strong predictive performance
- SHAP adds interpretability — both global (feature ranking) and local (individual reasons for prediction)

Threshold-Based Custom Logic:

- A custom 35% probability threshold was set to classify "Attrition Risk"
- This allows HR teams to interpret the model's predictions with practical decision-making in mind
- Helps balance risk identification without over-flagging employees

SHAP Interpretation In Action

- Global SHAP Summary: Top contributors visualized with mean SHAP values
- Local SHAP Explanations: Provided per prediction through the Streamlit app
- "Why this employee is at risk?" → Answered visually + via LLM-powered chatbot

RESULTS

Rank	Feature	Impact Type
1	MonthlyIncome	Negative (↓ risk with ↑ income)
2	OverTime	Positive (↑ risk when yes)
3	JobSatisfaction	Negative (↓ risk with ↑ satisfaction)
4	TotalWorkingYears	Negative (early career = higher risk)
5	DistanceFromHome	Positive (↑ risk with ↑ distance)

CONCLUSION

Key Takeaways:

- Employee attrition is predictable using machine learning when the right data is available.
- XGBoost, combined with SHAP explainability, achieved high accuracy (~88%) while remaining interpretable and actionable.
- The project highlights critical drivers of attrition like OverTime, Job Satisfaction, and Monthly Income — enabling HR to act early.

How the Solution Helps:

- For HR Teams: Identifies at-risk employees before they resign, enabling targeted retention strategies.
- For Decision Makers: Offers a transparent tool with visual explanations to support ethical and informed action.
- For Business Strategy: Aligns workforce stability with financial and operational goals.

FUTURE SCOPE

- Real-time HR data integration from internal systems (e.g., Workday, BambooHR)
- Historical dashboards to track attrition trends over time
- Automated reporting (PDF/CSV export with individual predictions)
- Multi-company generalization using transfer learning or federated data
- Ethical guardrails: Adding alerts when the model may show bias or overconfidence

LIVE APP DEMO

Streamlit Web App: (link)

- Interactive Prediction UI
- Users can adjust inputs like JobSatisfaction, OverTime, and MonthlyIncome using sliders, dropdowns, and numeric fields
- Upon clicking Predict Attrition Risk, the app returns:
  - A clear prediction (Yes/No)
  - Probability score
  - Risk classification using the 35% threshold logic
- SHAP-Based Explanations
  - Real-time feature impact for every prediction
  - Highlights “why” this particular employee might leave
  - Includes bar charts of feature contributions (local SHAP values)
- AI Assistant Chatbot
  - Integrated OpenAI-powered LLM chatbot
  - Answers only questions related to attrition predictions
  - Designed to support HR teams by explaining results in plain English

Q & A

- Thank You!
- I appreciate your time and attention throughout this presentation.
- I’m happy to answer any questions, discuss implementation ideas, or explore future enhancements.

Thank You

For your attention