Mini Project Report

on

Predicting Employee Attrition Using XGBoost and HR Analytics

Submitted By

Mohammed Saifuddin Tailor

**Course Name: Machine Learning**

****

**Department of Computer Science and Engineering**

(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

**Shri Ramdeobaba College of Engineering & Management, Nagpur 440013**

(An Autonomous Institute affiliated to Rashtrasant Tukdoji Maharaj Nagpur University Nagpur)

October 2025

**1. Problem Statement**

Employee attrition (turnover) is a critical concern for organizations as frequent departures result in high replacement costs, knowledge drain, and disruption to daily operations. For many organizations, the ability to proactively identify employees likely to leave enables focused retention strategies and reduces overall turnover costs.

This project addresses the following challenge:

* Can we predict whether an employee is likely to leave the organization, using historical demographic, performance, and work pattern data?
* By leveraging machine learning on the IBM HR Analytics dataset, the goal is to identify the key factors leading to attrition and enable data-driven decision-making for workforce management.

**2. Project Objectives**

1. **Attrition Risk Prediction:**  
   Develop an accurate predictive model to classify whether an employee will leave (attrition = Yes) or stay (attrition = No) using employee profile and job-related attributes.
2. **Key Drivers Analysis:**  
   Identify and interpret the most influential features affecting attrition within the organization.
3. **Business-Ready Deployment:**  
   Deploy an interactive web app using Streamlit for HR teams to easily input profiles and obtain automated risk predictions with actionable recommendations.
4. **Generalizability:**  
   Ensure the developed pipeline is consistent, easily extensible, and applicable to other similar HR datasets.

**3. Methodology**

**Data Source:**  
IBM HR Analytics Employee Attrition & Performance dataset from Kaggle.

**Workflow Steps:**

* 1. Data ingestion and EDA
  2. Data preprocessing (encoding, scaling, feature selection)
  3. Handling class imbalance using SMOTE
  4. Model selection and hyperparameter tuning
  5. Model evaluation and interpretation
  6. Pipeline creation and deployment

**Flowchart:**

Data Preprocessing

Raw Dataset

SMOTE Oversampling

Imbalanced Classes

RandomizedSearchCV for   
XGBoost Hyperparameters

Model Selection   
& Tuning

Model + Preprocessor  
Pipeline Saving

Model Evaluation   
& Interpretation

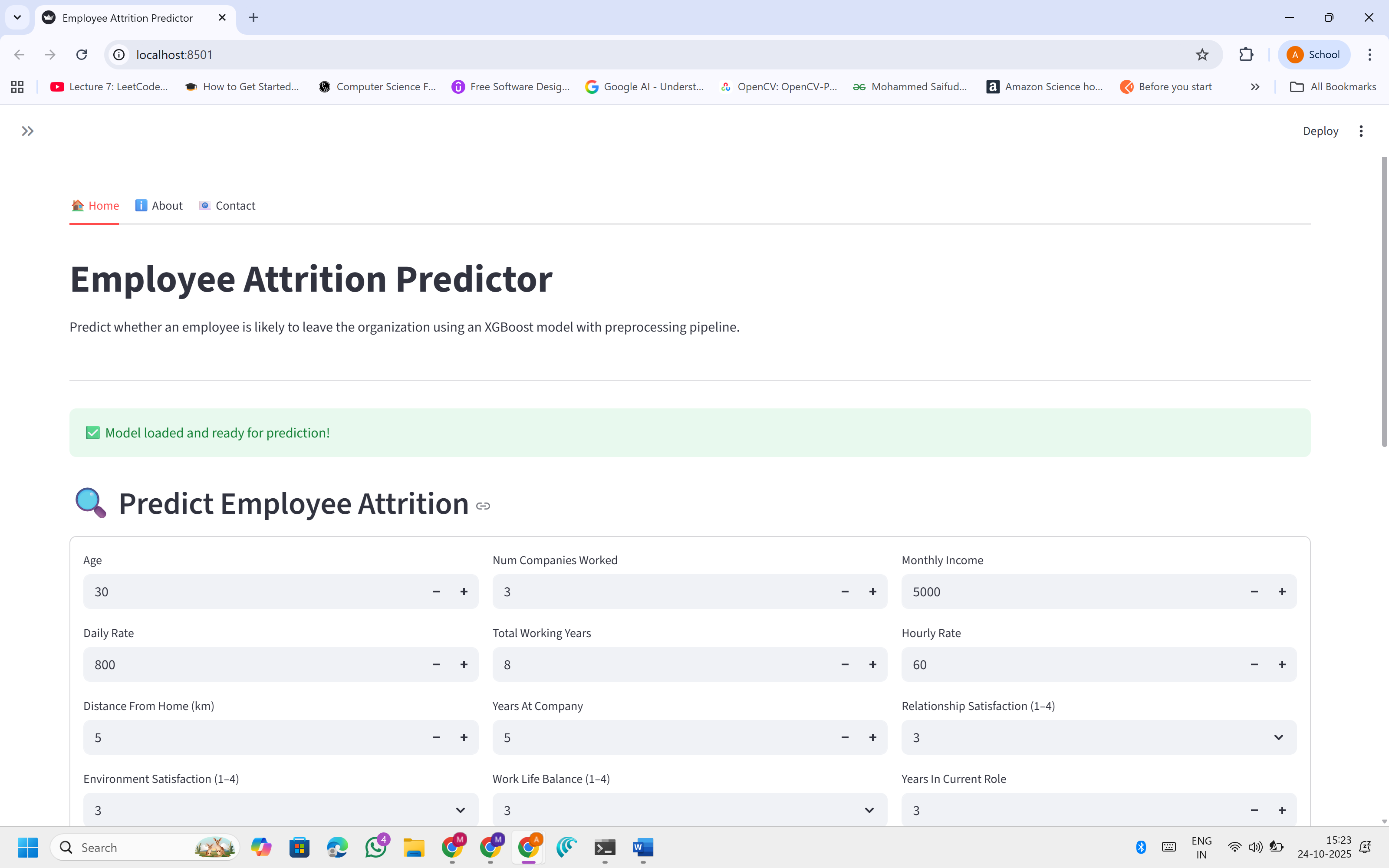
Streamlit Deployed Web App

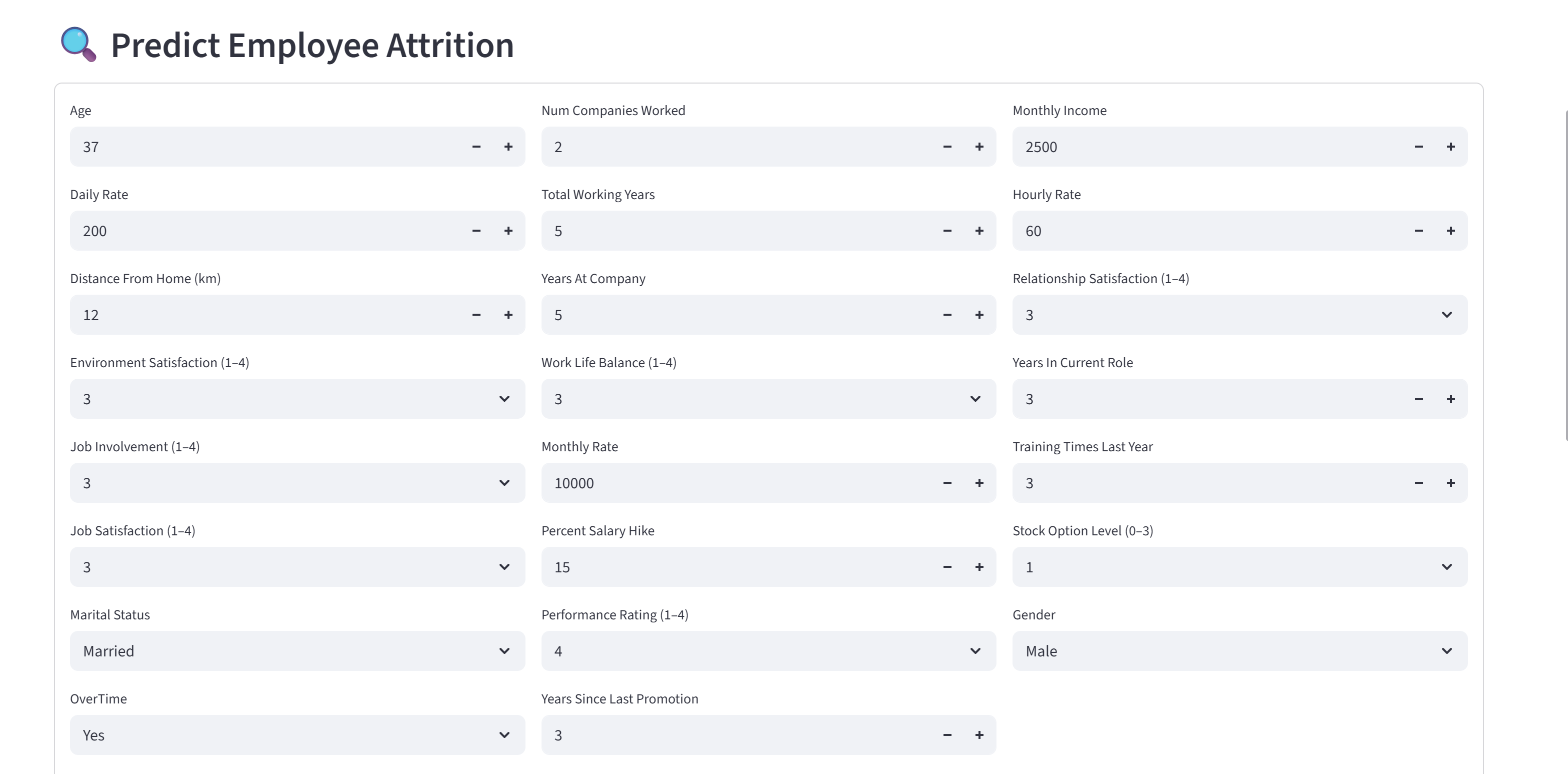
**4. Technology Stack**

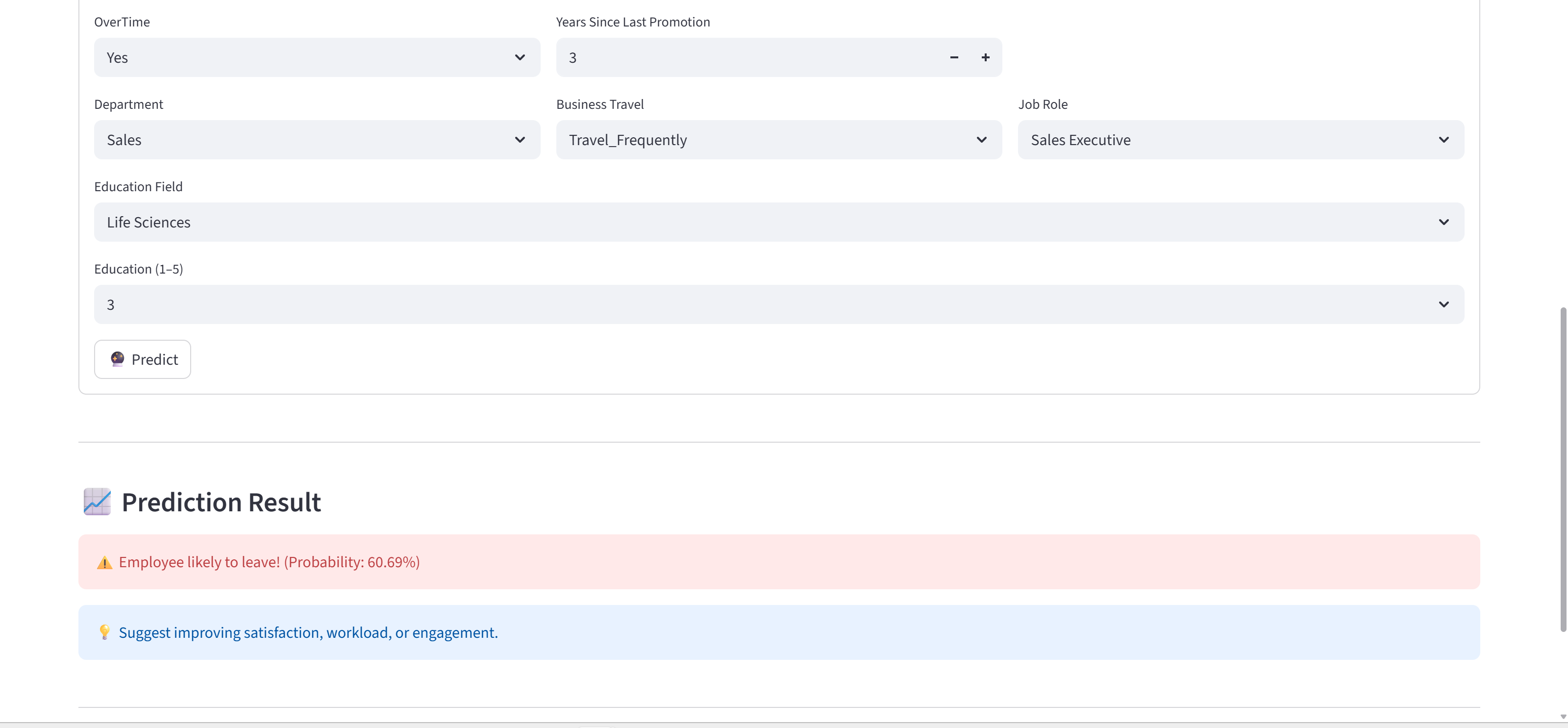
* Programming Language: Python 3.10 / 3.11
* **Core Libraries:**
  + pandas, numpy – Data manipulation
  + scikit-learn – ML preprocessing, metrics, model selection
  + imblearn – SMOTE
  + xgboost – Gradient boosting ML algorithm
  + streamlit – Web application framework
* **Visualization:**
  + matplotlib, seaborn
* **Model Saving:**
  + joblib – For pipeline serialization
* **Dataset:**
  + "IBM HR Analytics Employee Attrition & Performance" (Kaggle)
* **Deployment:**
  + Streamlit Cloud
  + GitHub (version control, CI/CD)

**5. Result**

* **Model Performance:**
  + Achieved test accuracy, ROC-AUC, and F1-score superior to baseline models (Logistic Regression, Random Forest, LightGBM).
  + XGBoost achieved an overall balanced F1-score and interpretability.
* **Interpretation:**
  + Key factors influencing attrition included: YearsAtCompany, OverTime status, MonthlyIncome, JobSatisfaction, and WorkLifeBalance.
* **Output:**
  + Successfully deployed as an interactive web app (EMPLOYEE ATTRITION PREDICTOR) on Streamlit Cloud.







**6. Conclusion**

This project demonstrates the development of a complete end-to-end pipeline for employee attrition prediction using machine learning techniques and cloud-based deployment. Leveraging the power of the XGBoost algorithm—coupled with data preprocessing and SMOTE-based oversampling—enabled the creation of a robust and interpretable predictive model. The final deployed solution provides HR professionals with a practical tool to assess employee attrition risk, interpret the main factors driving attrition, and take proactive and targeted retention actions. Importantly, the entire process is designed to be fully reproducible and scalable, ensuring that it can be easily adapted to other organizations or extended to different types of workforce datasets in the future.

**7. References**

1. Kaggle Dataset: **[IBM HR Analytics Employee Attrition & Performance](https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset" \t "_blank)**
2. XGBoost Documentation and Tutorials
3. Scikit-learn, imblearn, Streamlit official documentation