Employee Churn Prediction Project Documentation

1. Project Overview

This project aimed to predict employee churn (the likelihood that an employee will leave a company) using historical employee data. By accurately predicting churn, the model enables HR departments to proactively address potential issues, improve retention strategies, and reduce turnover costs.

The key components of this project included:

1. Data Collection and Preparation
2. Data Preprocessing
3. Exploratory Data Analysis
4. Feature Engineering
5. Model Selection and Training
6. Model Evaluation
7. Deployment and Predictions

2. Data Collection and Preparation

The dataset used for this project was **IBM HR Analytics Employee Attrition & Performance**, a well-structured dataset containing various employee attributes such as demographics, job role, performance, and satisfaction levels. It was loaded into a DataFrame, and unnecessary columns were removed or modified based on feature relevance for predicting churn.

Key Dataset Columns

- **Age, MonthlyIncome, DailyRate, TotalWorkingYears** - Numerical features related to employee experience and earnings.
- **JobLevel, PerformanceRating, JobSatisfaction** - Employee job satisfaction, performance, and level.
- **OverTime, BusinessTravel, Department** - Categorical features related to employee roles and work conditions.

3. Data Preprocessing

Data preprocessing included several steps to clean and transform the data, ensuring it was suitable for model training. Steps included:

- **Handling Missing Values:** Any missing values were filled or dropped based on their significance. However, the IBM dataset was clean and had minimal missing data.
- **Encoding Categorical Variables:** Categorical features (e.g., Department, JobRole, Gender) were one-hot encoded to ensure compatibility with machine learning models.
- **Scaling Numerical Features:** Numerical features were standardized using StandardScaler to normalize data and ensure model compatibility.

4. Exploratory Data Analysis (EDA)

EDA was conducted to understand relationships between features and the target variable (churn). Visualizations, including histograms, box plots, and correlation matrices, were created to explore:

- **Distribution of Key Features:** For example, income and age distributions.
- **Churn Analysis by Category:** Breakdown of churn rates by job role, department, and satisfaction level.

- **Correlations:** Identification of highly correlated features to manage feature redundancy.

5. Feature Engineering

Several feature engineering steps were taken to improve model input quality:

- **Binary Encoding for Attrition:** The target variable (Attrition) was converted to a binary format (0 for stay, 1 for leave).
- **One-Hot Encoding:** Applied to categorical variables (e.g., BusinessTravel, Department).
- **Derived Features:** Additional interaction features were considered (e.g., MonthlyIncome vs. JobSatisfaction) based on domain understanding.

6. Model Selection and Training

Chosen Models

Three models were selected to provide a balance between interpretability and accuracy:

1. **Logistic Regression:** Used as a baseline due to its simplicity and interpretability.
2. **Random Forest Classifier:** Offers feature importance insights and handles non-linear relationships.
3. **XGBoost Classifier:** A powerful gradient-boosting model known for high accuracy in classification tasks.

Training Process

Each model was trained on a train-test split of the data, with hyperparameters tuned to optimize performance. The train-test split ensured an unbiased evaluation on unseen data.

Hyperparameter Tuning

- **Random Forest:** Parameters such as n_estimators (number of trees) and max_depth were optimized using grid search.
- **XGBoost:** learning_rate, max_depth, and n_estimators were optimized using a similar approach for optimal performance.

7. Model Evaluation

Each model was evaluated based on multiple metrics:

- **Accuracy:** Proportion of correct predictions.
- **Precision and Recall:** To measure relevance of predictions and sensitivity to churn cases.
- **F1 Score:** A balanced metric, particularly useful in cases of class imbalance.
- **AUC-ROC Curve:** Evaluated models based on their area under the ROC curve for churn classification.

Evaluation Results

- **Random Forest Model:** Highest performance with an accuracy of ~85% and an AUC-ROC score of 0.90.
- **XGBoost Model:** Slightly higher performance than Random Forest but more computationally expensive.
- **Logistic Regression Model:** Lower performance but served as a useful baseline.

8. Deployment and Prediction

The final Random Forest model and scaler were saved using joblib, making them reusable for predictions on new data. A function was created to streamline employee churn predictions, enabling HR teams to input new employee data and receive churn probability outputs.

Example Prediction Workflow

Using the saved model and scaler, an employee's data (demographic, job role, and satisfaction levels) was entered to predict their probability of leaving. The model returned a probability score and classified the employee as likely to stay or leave.

9. Conclusion and Recommendations

This model allows HR departments to proactively address potential churn by identifying employees at high risk of leaving. Future steps include:

- **Further Model Tuning:** Experiment with additional models or fine-tune hyperparameters.
- **Incorporate Real-Time Data:** Integrate with HR databases to monitor employee status changes in real time.
- **Feedback Loop for Model Retraining:** Retrain the model periodically with updated data to maintain accuracy over time.