Title: Exploratory Data Analysis and Feature Engineering for Predictive Modeling

**Stakeholder:**

Name: IBM Corporation

Role: Data Science Team Lead

Problem: The stakeholder is interested in understanding and predicting employee turnover within the company. They want to identify factors that contribute to employee attrition and build a model that can predict the likelihood of an employee leaving the company.

**Dataset:**

Source: Kaggle - IBM HR Analytics Employee Attrition & Performance

**Description:** The dataset contains various features related to employee information, job satisfaction, work environment, and performance. The target variable is 'Attrition,' indicating whether an employee has left the company or not.

**Problem Definition:**

Question to Answer: What factors contribute to employee attrition, and can we predict the likelihood of an employee leaving?

**Target Variable:** 'Attrition' (Yes/No)

**Features:**

Personal Information: Age, Gender, Marital Status

Job-related: Department, Job Role, Years at Company, etc.

Performance: Job Satisfaction, Work-Life Balance, etc.

**Data Exploration:**

**Distribution:** Check the distribution of the target variable 'Attrition' to understand the class balance.

**Missing Values:** Identify and handle missing values in the dataset.

**Outliers:** Explore numerical features for outliers and decide whether to handle or keep them.

**Feature Engineering Techniques:**

Handling Missing Values:

Method: Impute missing values using mean, median, or mode based on the nature of the feature.

**Pros and Cons:** This ensures that the model can still use the information from other features. However, it may introduce bias if missing values are not missing completely at random.

**Outlier Detection:**

**Method**: Use statistical methods (like IQR) to detect and handle outliers in numerical features.

**Pros and Cons**: Helps in maintaining the integrity of the data, but removing outliers might lead to loss of information.

**Encoding Categorical Variables:**

**Method:** Use one-hot encoding for categorical variables like 'Department' and 'Job Role.'

Pros and Cons: Ensures the model can understand categorical features, but may increase the dimensionality of the dataset.

Feature Scaling:

Method: Apply Min-Max scaling to numerical features to bring them to a similar scale.

Pros and Cons: Helps in preventing features with larger scales from dominating the model. However, it might be sensitive to outliers.

Code Implementation:

Link to GitHub Repository: Employee\_Attrition\_Prediction

Code Files:

data\_exploration.ipynb: Data exploration and visualization.

feature\_engineering.py: Python script for feature engineering functions.

attrition\_prediction\_model.ipynb: If modeling is performed.

Notes on Deployment:

New Values: Handle new values by ensuring the same preprocessing steps are applied before making predictions.

Missing Data: For deployment, if new data has missing values, use the imputation strategy identified during the exploration phase.

Outliers: Decide during deployment whether to clip or transform outliers based on the analysis performed during exploration.

Conclusion:

This report provides a comprehensive exploration of the dataset, outlines the problem, and discusses feature engineering techniques applied to prepare the dataset for predictive modeling. The code is structured and documented for reproducibility and readability.