Building a Machine Learning Model for Predicting Employee Turnover

As a machine learning specialist, you have been hired by a large corporation to help them analyze their employee data in order to better understand their workforce and make informed decisions related to employee retention, productivity, and engagement.

The company has provided you with a dataset containing information on employee Education, JoiningYear, City, PaymentTier, Age, Gender, EverBenched, ExperienceInCurrentDomain, and LeaveOrNot.

The employee dataset includes the following features:

• **Education:** The level of education attained by the employee (e.g., high school, bachelor's degree, master's degree, etc.).

• **JoiningYear:** The year in which the employee joined the company. • City: The city in which the employee is based.

• **PaymentTier:** The payment tier or salary bracket for the employee (e.g., entry-level, mid-level, senior-level, etc.).

• **Age:** The age of the employee. • Gender: The gender of the employee.

• **EverBenched:** A binary indicator of whether the employee has ever been benched (i.e., temporarily removed from a project or assignment).

• **ExperienceInCurrentDomain:** The number of years of experience the employee has in their current field or domain.

• **LeaveOrNot:** A binary indicator of whether the employee has left the company. Your goal is to develop a machine learning model using logistic regression that can accurately predict whether an employee will leave or not (employee turnover) and identify the key drivers of employee engagement and productivity.

The goal of the machine learning project is to develop a logistic regression model that can accurately predict whether an employee will leave the company or not based on the other variables in the dataset. In addition to developing the model, exploratory data analysis, data processing, data preprocessing, feature engineering and selection, modeling, and model optimization will be performed.

Therefore, in addition to developing the logistic regression model, you should also conduct exploratory data analysis to identify any patterns or trends in the data that may not be immediately apparent. You should use visualizations and statistical techniques to highlight important relationships and correlations in the data, and use this information to inform your model-building process.

Finally, perform model selection using cross-validation to determine which of the models (SVM, Random Forest, Gradient Boosting, KNN, & Naïve Bayes) performs best.