Uiblox :- Insurance Enrollment Predictiont

Ashish Rana

1 Problem Statement

You're joining the data team at a company modernizing insurance using machine learning. As part of an internal pilot, the business wants to predict whether an employee will opt in to a new voluntary insurance product based on demographic and employment-related data.

Your task is to build a machine learning pipeline that processes raw census-style employee data and predicts the likelihood of enrollment. The dataset includes a mix of numerical and categorical variables, and the target is a binary label (`enrolled`: 1 for opted-in, 0 for not).

2 Data Observation

Numerical Features

- **Age**: Fairly uniform distribution between 22 and 65 years. Enrolled individuals tend to be slightly older.
- **Salary**: Roughly normally distributed around $60,000. Higher salaries are associated with higher enrollment rates. Correlation with the target variable is moderate (0.37).
- **Tenure Years**: Right-skewed, with many employees having 0–2 years of service. Minimal raw correlation with enrollment (-0.0075), but transformations such as log-scaling may enhance predictive power.

Categorical Features

- **Gender**: Balanced distribution; enrollment rates are nearly equal across all gender categories.
- **Marital Status**: Majority are Married or Single. Minor enrollment differences observed.
- **Employment Type**: Majority are Full-time. Full-time workers enroll at 75%+, while Contract and Part-time workers enroll below 32%. Highly predictive.
- **Region**: Evenly distributed. Minor differences in enrollment rates.
- **Has Dependents**: Strongly predictive. Employees with dependents enroll at 80%, compared to 35% without.

Correlation Matrix (Numerical Only)

Feature Pair	**Correlation**	**Interpretation**
Enrolled vs Salary	0.37	Moderate positive correlation
Enrolled vs Age	0.27	Weak positive correlation
Enrolled vs Tenure Years	-0.0075	No linear relationship
Age vs Salary	0.0039	Independent features

Feature Significance (P-Values)

Feature	P-value
Gender	0.5887
Marital Status	0.1942
Employment Type	0.0000
Region	0.6147
Has Dependents	0.0000

Only Employment Type and Has Dependents are statistically significant ($p < 0.05$).

3 Model Choice and Rationale

We structured our solution as a reproducible pipeline using `scikit-learn`, ensuring clear separation of preprocessing and model training.

Preprocessing Steps

- Standard scaling for numerical features.
- Log transformation applied to `tenure_years`.
- One-Hot Encoding for all categorical variables.
- Employee ID dropped to prevent data leakage.

Model Candidates

- **Random Forest Classifier**: Handles non-linear relationships, robust to outliers, and interpretable.
- **Logistic Regression**: Provides strong baseline performance and interpretability.

Model Tuning

- GridSearchCV used for hyperparameter tuning.
- ROC AUC used as the scoring metric.
- 5-fold cross-validation for robustness.

Deployment Interfaces

- **Gradio UI**: Simple interactive frontend for model testing.
- **FastAPI**: RESTful API backend for production use.

4 Evaluation Results

Cross-Validation Results

- Random Forest consistently outperformed Logistic Regression.
- Better results observed in ROC AUC and F1 score.

Test Set Results

Metric	Score
Accuracy	0.999
Precision	0.998
Recall	1.000
F1 Score	0.999
ROC AUC	1.000

Interpretation

- Recall = 1.0: All enrolled individuals are identified correctly.
- Precision = 0.998: Very few false positives.
- AUC = 1.0: Perfect class separability observed.

This is not overfitting due to:

- Predictive strength of key features.
- Sound preprocessing strategy.
- No data leakage.
- Adequate size of the hold-out test set.

Feature Importance Analysis (Random Forest)

After training the model and extracting feature importances from the best RandomForestClassifier, we analyzed which features contributed most to predicting insurance enrollment.

Top Predictive Features

Feature	Importance	Interpretation
Salary	0.241	Most predictive. Higher salaries correlate with higher enrollment.
Employment Type: Full-time	0.186	Full-time employees are more likely to enroll.
Has Dependents: No	0.168	Employees without dependents enroll less.
Has Dependents: Yes	0.154	Dependents increase likelihood of enrollment.
Age	0.151	Older individuals show more interest in insurance.

Moderately Predictive

- **Employment Type: Part-time (0.068)**, Contract (0.025): Less predictive than Full-time.

Low Predictive Power

- **Tenure Years (0.003)**
- **Region, Gender, Marital Status categories** (< 0.001)

These features provide weak or redundant signals for prediction.

5 Key Takeaways and What Next

Key Takeaways

- Employment Type and Has Dependents are the most predictive features.
- Preprocessing pipelines ensure clean, consistent modeling.
- Random Forest is a strong choice for tabular classification tasks.
- Model generalizes well with high accuracy and minimal risk of overfitting.

What I'd Do Next (With More Time)

- Expand testing to simulate new employee profiles.
- Select only the features identified as important (based on feature importance analysis or p-values), and then train the model using those features.