

Project 5 : Advanced Attrition Models (Tuned Random Forest + XGBoost)

In this notebook, we benchmark and enhance predictive models for employee attrition.

- Start with tuned **Random Forest**
- Introduce **XGBoost** (gradient boosting)
- Compare performance vs Logistic Regression

Goal:

Improve predictive accuracy while balancing interpretability and business storytelling.

1 Load & Preprocess Data

Dataset shape: (1470, 44)

2 Train-Test Split & Scaling

- Logistic Regression → needs scaling
- RF & XGBoost → raw features

3 Random Forest Hyperparameter Tuning

Best Params (RF): {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 100}

Best CV Accuracy (RF): 0.8665055896141363

Out[7]: ['models/random_forest_tuned.pkl']

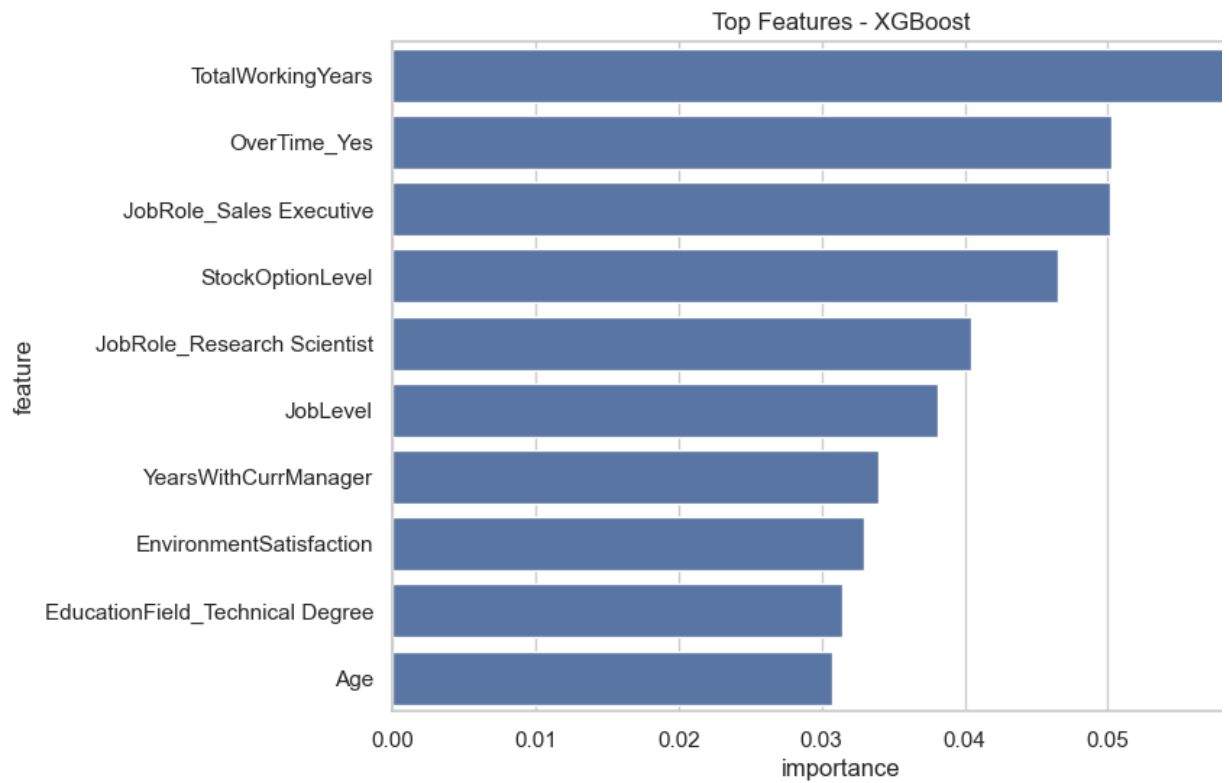
4 Train XGBoost Model

XGBoost Accuracy: 0.8639455782312925

XGBoost ROC AUC: 0.7741407528641572

Out[9]: ['models/xgboost_attrition_model.pkl']

5 Feature Importance



6 Model Comparison

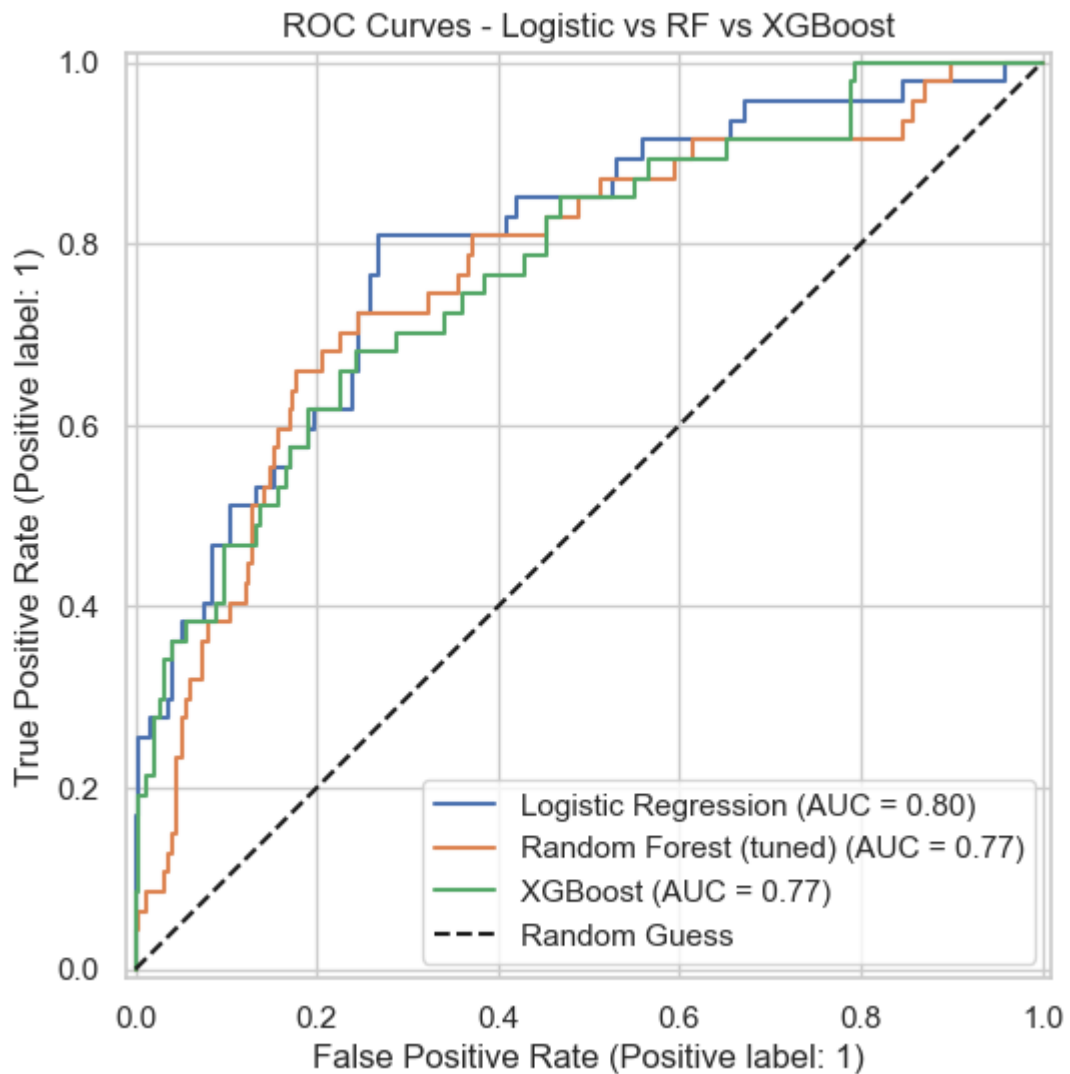
	Model	Accuracy	ROC AUC
0	Logistic Regression	0.751701	0.798260
1	Random Forest (tuned)	0.836735	0.768886
2	XGBoost	0.863946	0.774141



ROC Curves for All Models

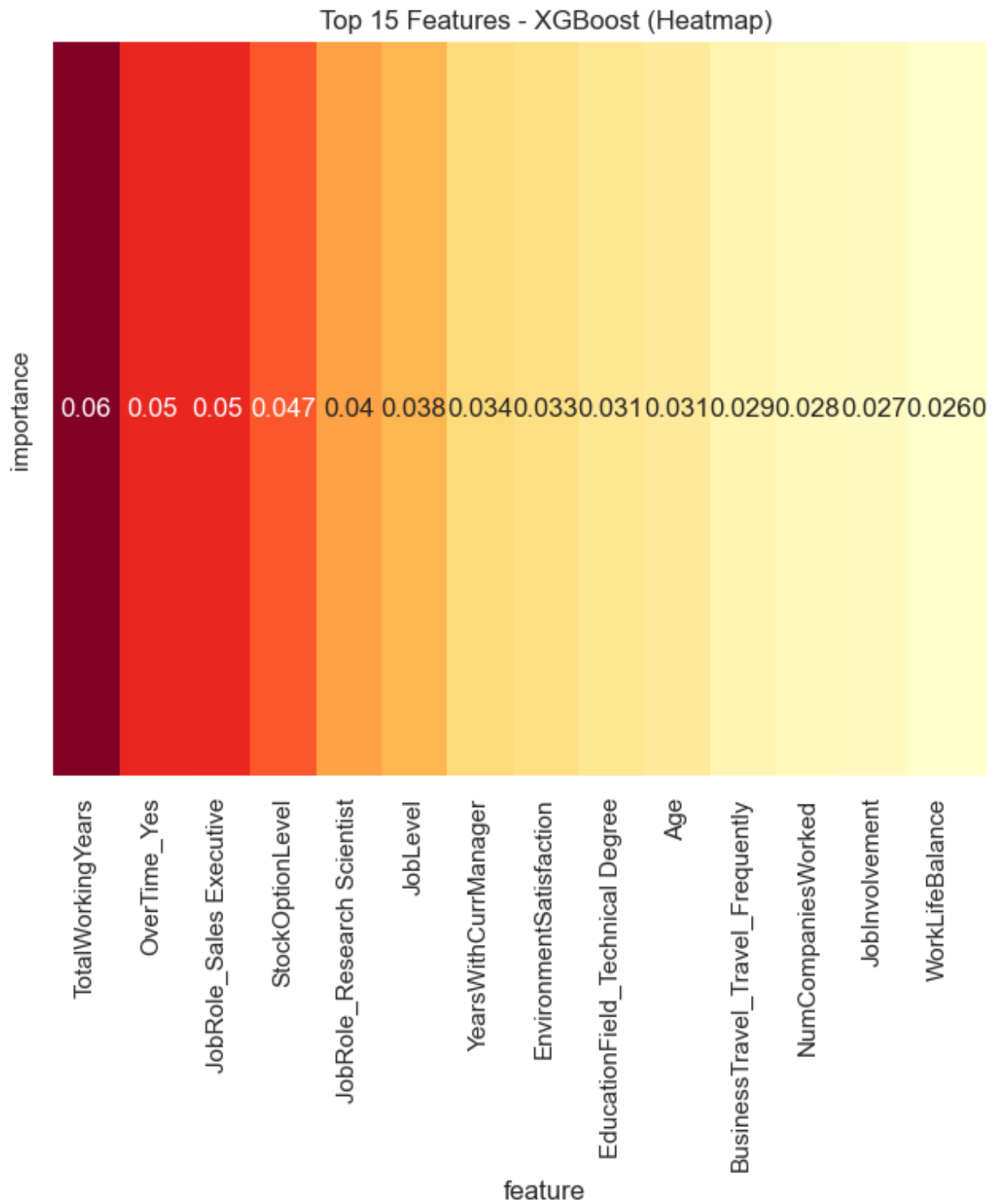


Shows how Logistic, RF, and XGBoost trade off sensitivity vs specificity — RF edges ahead !



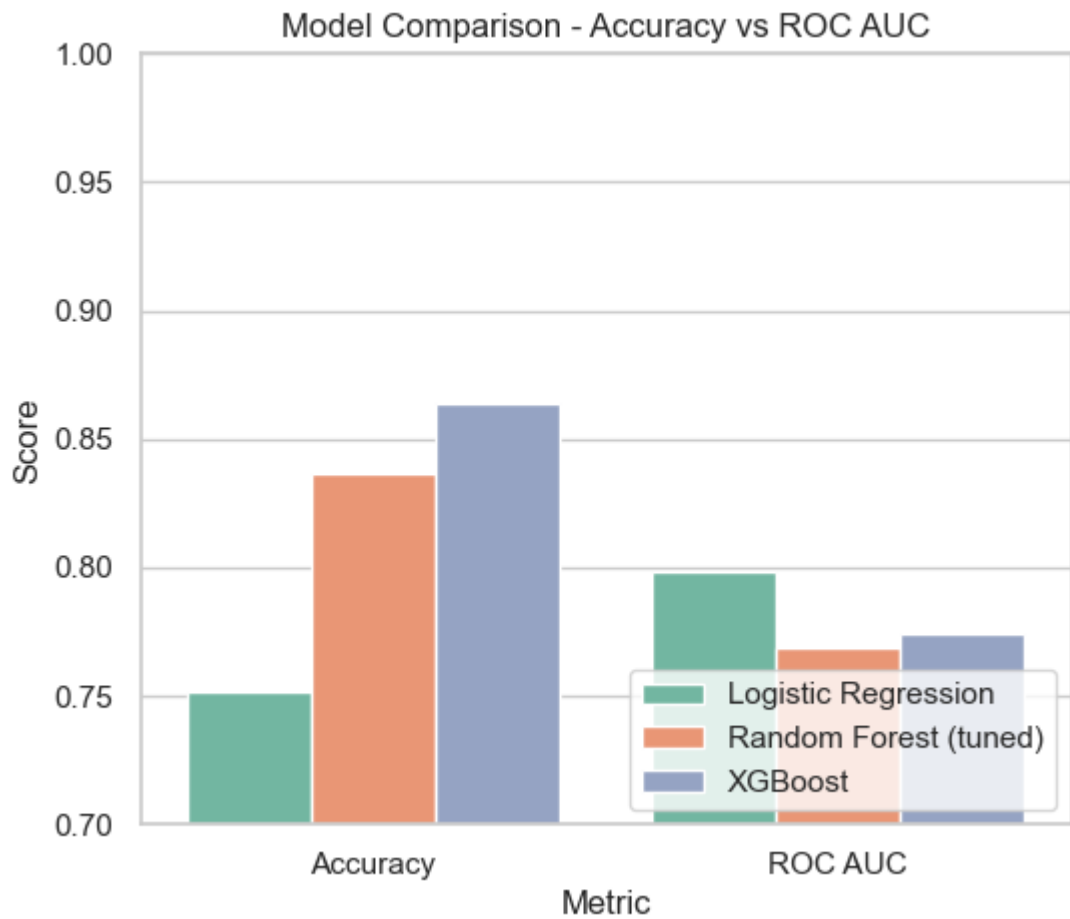
Feature Importances Heatmap (XGBoost)

🔥 Top 1 5 features by importance — Total Working Years, Overtime, and Job Level- Sale Executive are top 3 attrition risk signals.



Model Comparison Barplot

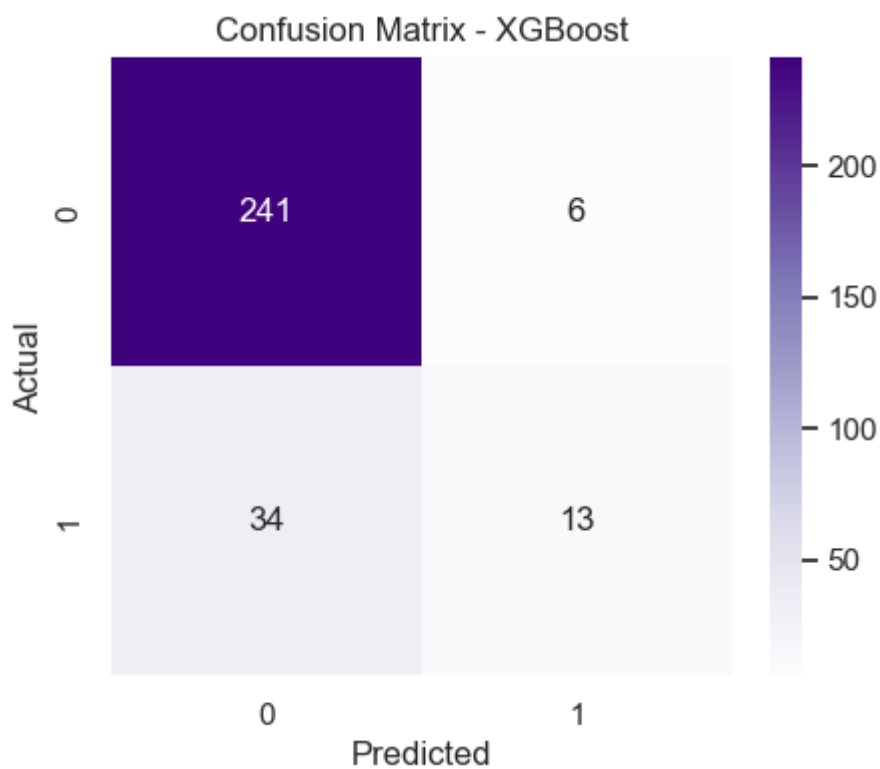
Quick side-by-side accuracy & ROC AUC.



Confusion Matrix Heatmap (XGBoost only)



Visual breakdown of predictions — where XGBoost gets it right (and where it misses).





Conclusions (Project 5)

- **Logistic Regression**

Accuracy ~ 75 %, ROC AUC ~ 0.79 → simple & interpretable

- **Tuned Random Forest**

Accuracy ~ 83 %, ROC AUC ~ 0.76 → improved performance, good at non-linear signals

- **XGBoost**

Accuracy ~ 86 %, ROC AUC ~ 0.77 → strongest predictive performance

Key Insights:

- OverTime and Sales roles remain consistent predictors across models
- Random Forest & XGBoost highlight additional signals like MonthlyIncome and Age buckets
- Logistic is best for storytelling, XGBoost for prediction