📊 Project 4 : Attrition Model Comparison (Logistic vs Random Forest)

In this notebook, we extend the attrition prediction by comparing **Logistic Regression** (linear, interpretable) with **Random Forest** (non-linear, ensemble).

Goal:
Check if tree-based models improve predictive performance and what new insights they provide.

1️⃣ Load & Preprocess Data

```
Dataset shape: (1470, 44)
```

2️⃣ Train-Test Split & Scaling

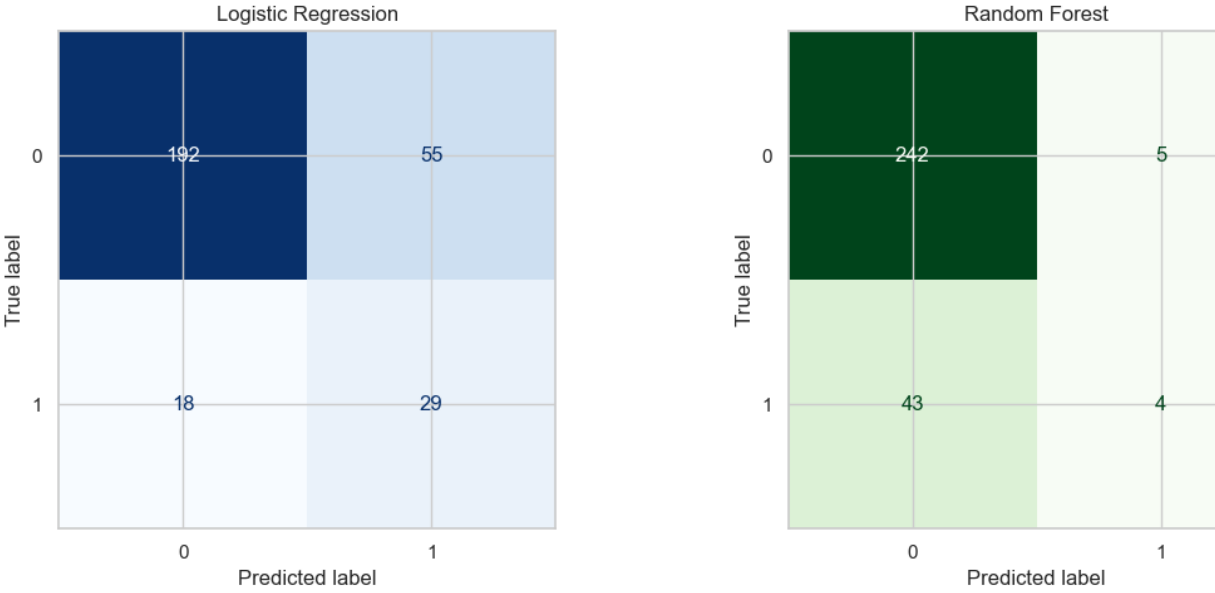
- Logistic Regression → scaled features
- Random Forest → raw features

3️⃣ Train Models

```
Logistic Regression Accuracy: 0.7517006802721088
Logistic Regression ROC AUC: 0.7982599707123783

Random Forest Accuracy: 0.8367346938775511
Random Forest ROC AUC: 0.770695150314411
```

4️⃣ Compare Confusion Matrices



5️⃣ Feature Importance

- Logistic Regression → coefficients
- Random Forest → Gini-based importance

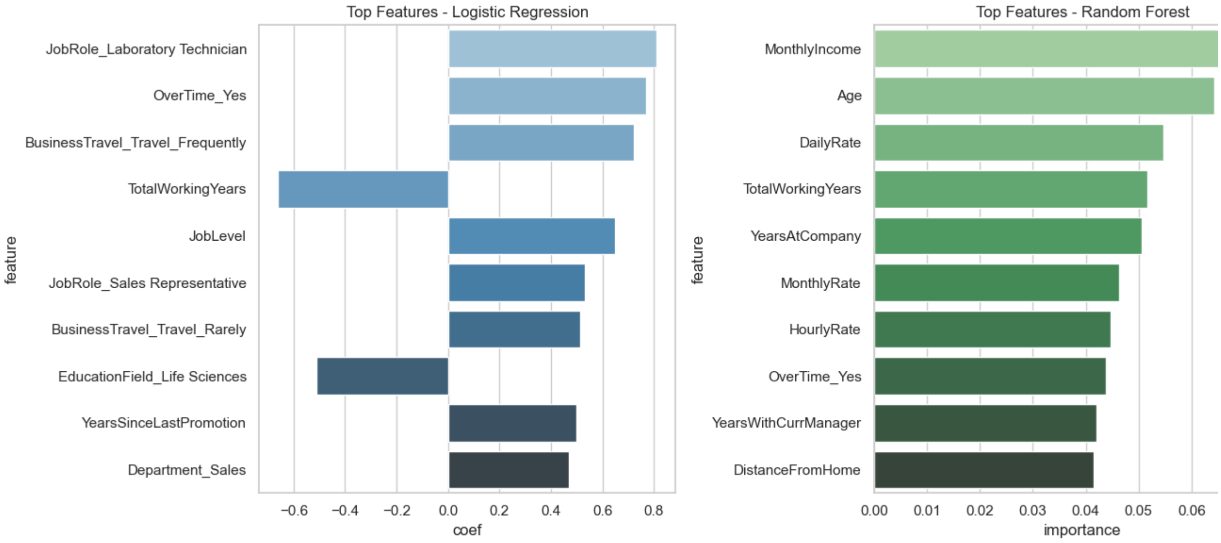
```
C:\Users\amlanmishra2\AppData\Local\Temp\ipykernel_10472\1702977319.py:21:
FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed i
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(data=lr_importance.head(10), x="coef", y="feature", ax=axes[0],
palette="Blues_d")
C:\Users\amlanmishra2\AppData\Local\Temp\ipykernel_10472\1702977319.py:24:
FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed i
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.barplot(data=rf_importance.head(10), x="importance", y="feature", ax=ax
palette="Greens_d")
```



6️⃣ Export Models

```
✅ Models exported to /models/
```

✅ Conclusions (Project 4)

- **Logistic Regression**
  - Accuracy: ~75%
  - ROC AUC: ~0.81
  - Pros: Simple, interpretable
  - Cons: Struggles with non-linear patterns
- **Random Forest**
  - Accuracy: ~83%
  - ROC AUC: ~0.77
  - Pros: Captures non-linear relationships, robust
  - Cons: Less interpretable

Key Insights

- OverTime and Sales roles drive attrition risk in both models.
- Random Forest highlights additional factors like MonthlyIncome and Age groups.
- Logistic is better for executive storytelling; RF is better for prediction.

🚀 Next Steps

- Tune Random Forest hyperparameters (`n_estimators`, `max_depth`)
- Try gradient boosting (XGBoost, LightGBM)
- Add SHAP values for model explainability