Lindsey Chenault
Discussion

The Random Forest model has the most accurate ROC curve due to a test accuracy of 91.02% and a strong AUC, which signifies exceptional discrimination (important for identifying defaulters). Therefore, I recommend using Random Forest to predict loan default probability. Gradient Boosting is second with an accuracy of 89.93%, additionally displaying strong performance but with a slightly lower AUC. Hence, Random Forest is still the recommended model. In contrast, Gradient Boosting earns the lowest RMSE of 2,145.84 for predicting loss amount, which makes it the most precise model for estimating expected losses. Random Forest is second, while Linear Regression models underperform because of linear assumptions, and the Decision Tree functions worst with an RMSE of 4,060.58.

TARGET_BAD_FLAG

Tree-based models surpass Logistic Regression:

- **Random Forest**: 91.02%
- **Gradient Boosting**: 89.93%
- **Logistic Regression models**: 86.99%–88.26%

Random Forest displays slight overfitting (training accuracy is 99.92%), but its predictive power explains its usage. Logistic models are more interpretable and stable across testing, though less flexible for non-linear relationships. Feature selection is adequate (REG_STEPWISE almost matches the whole model's accuracy, seizing pivotal signals with fewer variables.

TARGET_LOSS_AMT

- **Gradient Boosting**: RMSE 2,145.84
- **Random Forest**: RMSE 2,447.50
- **Linear Regression (REG_ALL)**: RMSE 3,072.29
- **Stepwise/Tree-based Reduced Models**: RMSE 3,485.72
- **Decision Tree**: RMSE 4,060.58

Gradient Boosting excels at grasping complex non-linear relationships. It should be the production model for loss prediction, especially given the importance of precise loss estimation for capital planning.

Coefficient Interpretations

For Stepwise Logistic Regression, positive coefficients on TRUNC_M_DEBTINC, O_M_VALUE, O_IMP_DELINQ, TRUNC_IMP_DELINQ, and TRUNC_IMP_DEBTINC align with default risk factors like high debt, delinquencies, and extreme values. I recommend checking X_train[[TRUNC_M_DEROG, TRUNC_IMP_DELINQ]].corr() for collinearity. If high, consider removing or combining these.

In the linear regression model, the fundamental drivers of increased loss are TRUNC_M_DEBTINC, TRUNC_IMP_CLNO, TRUNC_IMP_DEBTINC, and TRUNC_LOAN. A longer credit age (TRUNC_IMP_CLAGE) lowers expected losses, supporting its role as a proxy for borrower reliability.

Final Recommendations

- **Classification (Default Prediction)**:
    - **Best**: Random Forest (91.02%)
    - **Next Best**: Gradient Boosting (89.93%)

- **Regression (Loss Prediction)**:
    - **Best**: Gradient Boosting (RMSE 2,145.84)
    - **Next Best**: Random Forest (2,447.50)
    - **Linear Benchmark**: REG_ALL (3,072.29)

For deployment:

- Use Random Forest for default prediction where accuracy is critical.
- Use REG_STEPWISE when interpretability is needed (e.g., for compliance).
- Use Gradient Boosting to predict loss amounts.

Overall, ensemble models transcend traditional methods, delivering robust solutions for the classification of credit risk and estimating financial loss.