DACANAY | RAMILO

PREDICTING CUSTOMER CHURN WITH REGRESSION-BASED AND TREE-BASED METHODS

CAPSTONE PROJECT

20 May 2025

INTRODUCTION

- EXPLORED TELECOM CUSTOMER CHURN USING EDA AND MACHINE LEARNING.
- **GOAL:** PREDICT CHURN BASED ON USAGE METRICS AND CUSTOMER BEHAVIOR.
- EVALUATED MODEL PERFORMANCE ON BOTH TRAINING AND TEST SETS.

METHODOLOGY

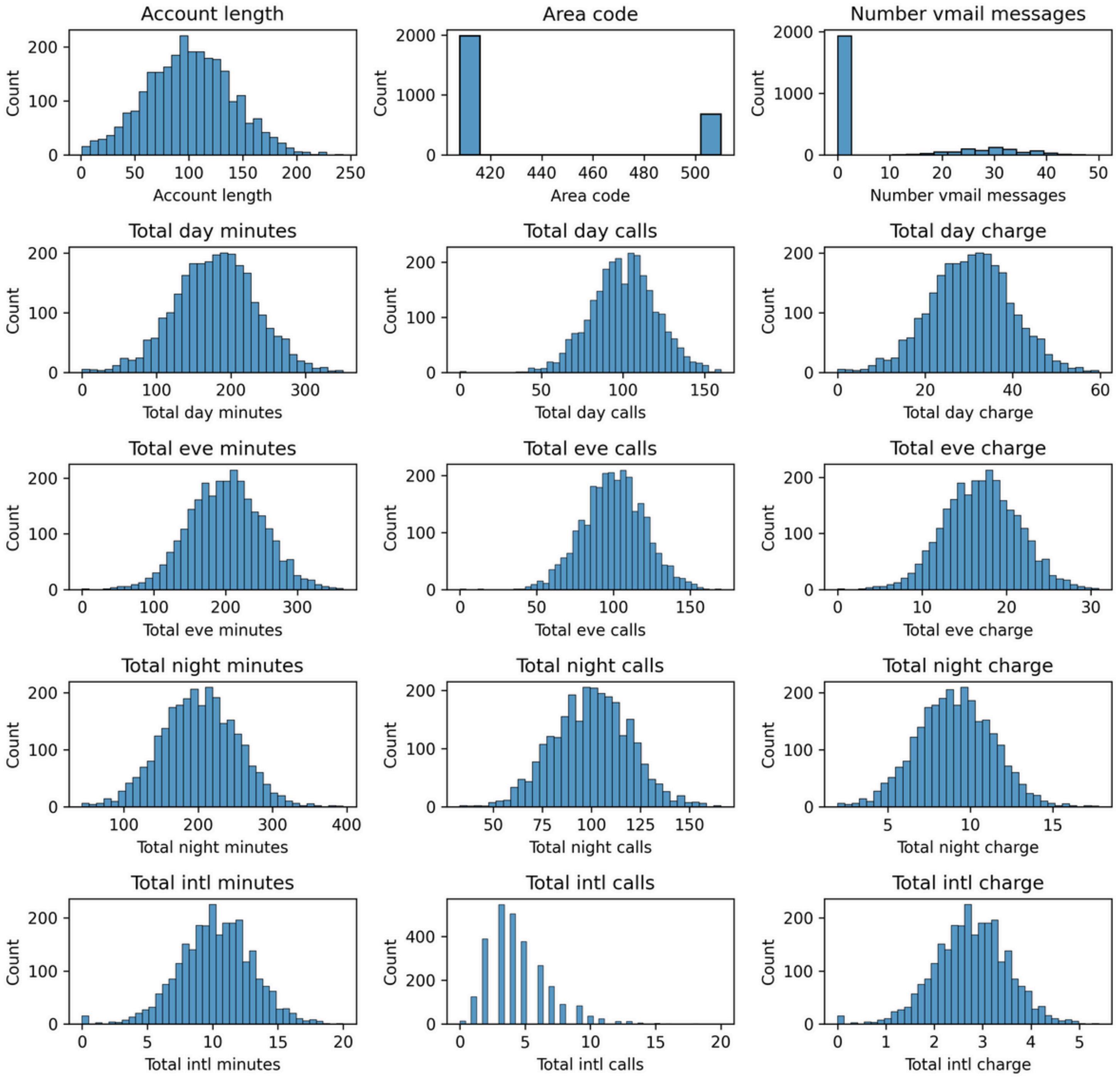
DATA

ORANGE TELECOM’S CHURN DATASET

Variable Name	Type	Variable Name	Type	Variable Name	Type	Variable Name	Type
State	Categorical	Total eve calls	Numerical	Number vmail messages	Numerical	Total intl minutes	Numerical
Account length	Numerical	Total eve charge	Numerical	Total day minutes	Numerical	Total intl calls	Numerical
Area code	Numerical	Total night minutes	Numerical	Total day calls	Numerical	Total intl charge	Numerical
International plan	Categorical	Total night calls	Numerical	Total day charge	Numerical	Customer service calls	Numerical
Voice mail plan	Categorical	Total night charge	Numerical	Total eve minutes	Numerical	Churn	Categorical

- Two datasets are provided: churn-80 and churn-20.
- Data split: churn_80 (training), churn_20 (testing).

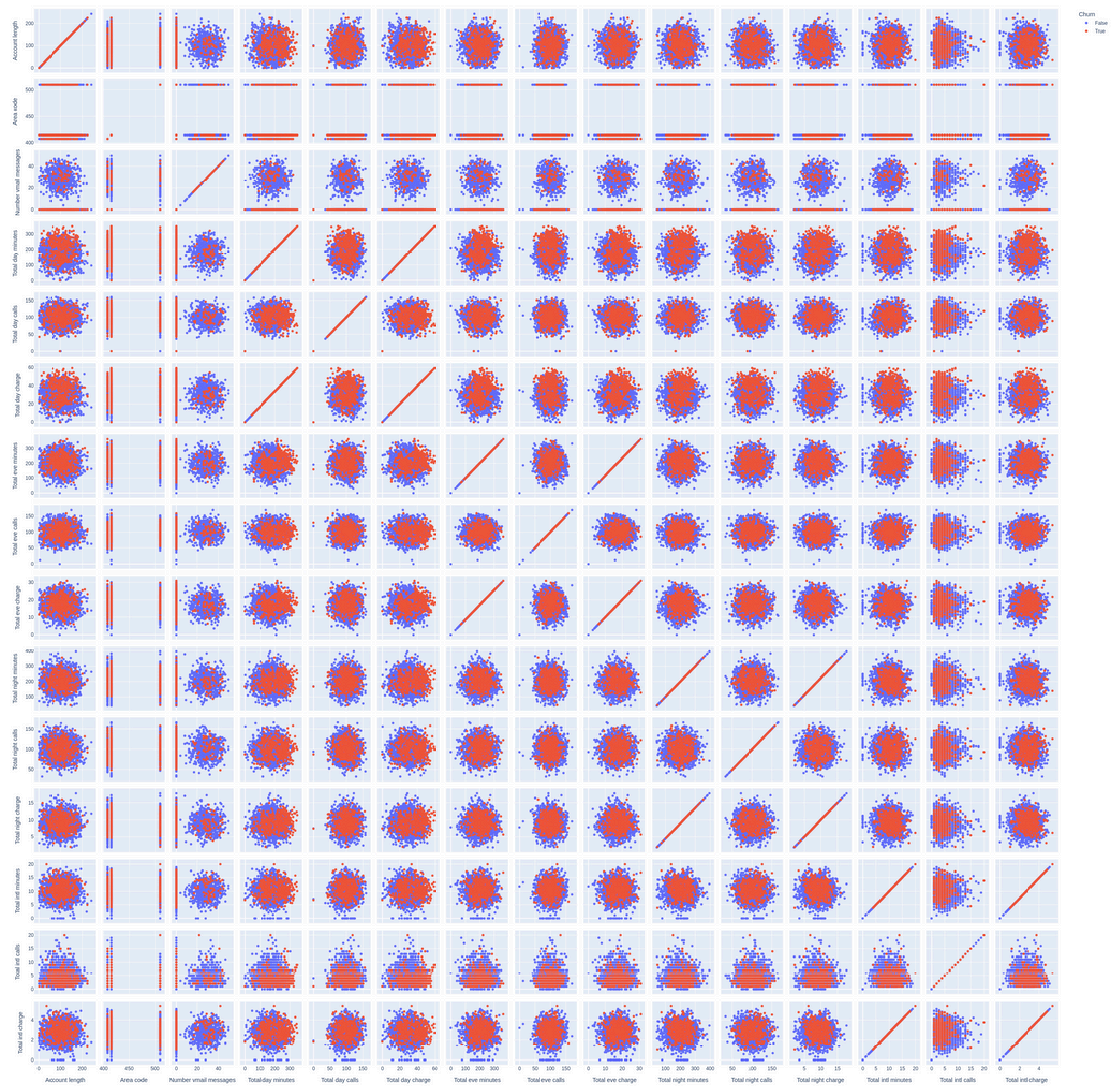
EXPLORATORY DATA ANALYSIS



UNIVARIATE ANALYSIS

- Does not exhibit severe outliers or anomalies.
- Most variables demonstrate distributional properties that are favorable for modeling.
- Skewed variables like voicemail usage and international calls may benefit from transformation or stratification to enhance model performance.
- Overall, the distributions suggest that the data is well-behaved and suitable for statistical modeling and machine learning applications.

EXPLORATORY DATA ANALYSIS



BIVARIATE ANALYSIS

The scatter matrix confirms the presence of highly correlated redundant variables, limited linear separability between churn classes, and the necessity for nonlinear classification models and feature engineering to improve churn prediction.

PREPROCESSING STEPS

- TYPE CONVERSION & ENCODING
- ONE-HOT ENCODING & YEO-JOHNSON TRANSFORMATION
- STANDARDIZATION
- ADDRESSED IMBALANCE WITH RANDOM UNDER-SAMPLING + SMOTE.

MODEL TRAINING

- **MODELS USED**
  - DECISION TREE
  - RANDOM FOREST
  - GRADIENT BOOSTING
  - LOGISTIC REGRESSION (WITH GRID SEARCH TUNING)
- EVALUATED USING ACCURACY, PRECISION, RECALL, F1-SCORE, ROC AUC.

RESULTS

Model Name	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Decision Tree	0.797601	0.39899	0.831579	0.539249	0.811768
Random Forest	0.866567	0.52	0.821053	0.636735	0.847589
Gradient Boosting	0.878561	0.546667	0.863158	0.669388	0.872138
Logistic Regression	0.7991	0.397906	0.8	0.531469	0.799476
Grid Search	0.806597	0.406593	0.778947	0.534296	0.795068

- **Gradient Boosting**: Best overall metrics (Recall, F1-Score, ROC AUC).
- **Random Forest**: Strong generalization.
- **Decision Tree**: High recall, low precision → overpredicts churn.
- **Logistic Regression**: Lower scores, but interpretable.

FEATURE IMPORTANCE

- **Top features:** Customer service calls, usage minutes, intl calls.
- Tree-based models emphasize sharp splits.
- Linear models spread importance more evenly.
- Feature patterns are consistent across models.

CONCLUSION

- Gradient Boosting offers optimal balance of performance.
- Behavioral metrics are key churn predictors.
- Logistic Regression remains valuable for interpretation.
- Proper preprocessing and resampling are crucial.