Predicting Customer Churn Using Machine Learning

Laura Greenwald
Mentor: Ernest Selvaraj

Executive Summary

Project Goal & Approach

Developing a predictive model using data from **7,043 customers** to identify at-risk individuals and reduce revenue loss.

Key Result

A tuned **Gradient Boosting** classifier achieved an **86.5%** ROC-AUC score.

The Business Problem

Context

Customer churn is a major challenge in telecommunications; retaining existing customers is significantly more cost-effective than acquiring new ones.

Objectives

- Identify high-risk customers.
- Understand key churn drivers.
- Develop targeted retention strategies.

THE DATA

21 FEATURES

- customerID
- gender
- SeniorCitizen
- Partner
- Dependents
- tenure
- PhoneService
- MultipleLines
- InternetService
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV
- StreamingMovies
- Contract
- PaperlessBilling
- PaymentMethod
- MonthlyCharges
- TotalCharges
- Churn

CUSTOMER DATA POINTS

7,043

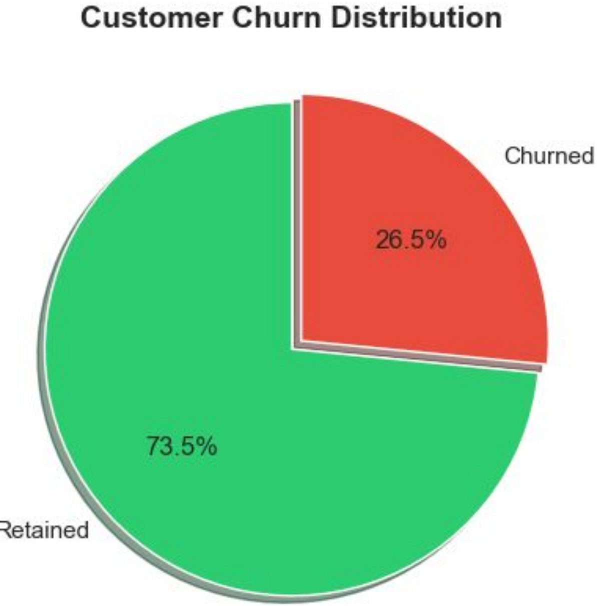
CUSTOMERS

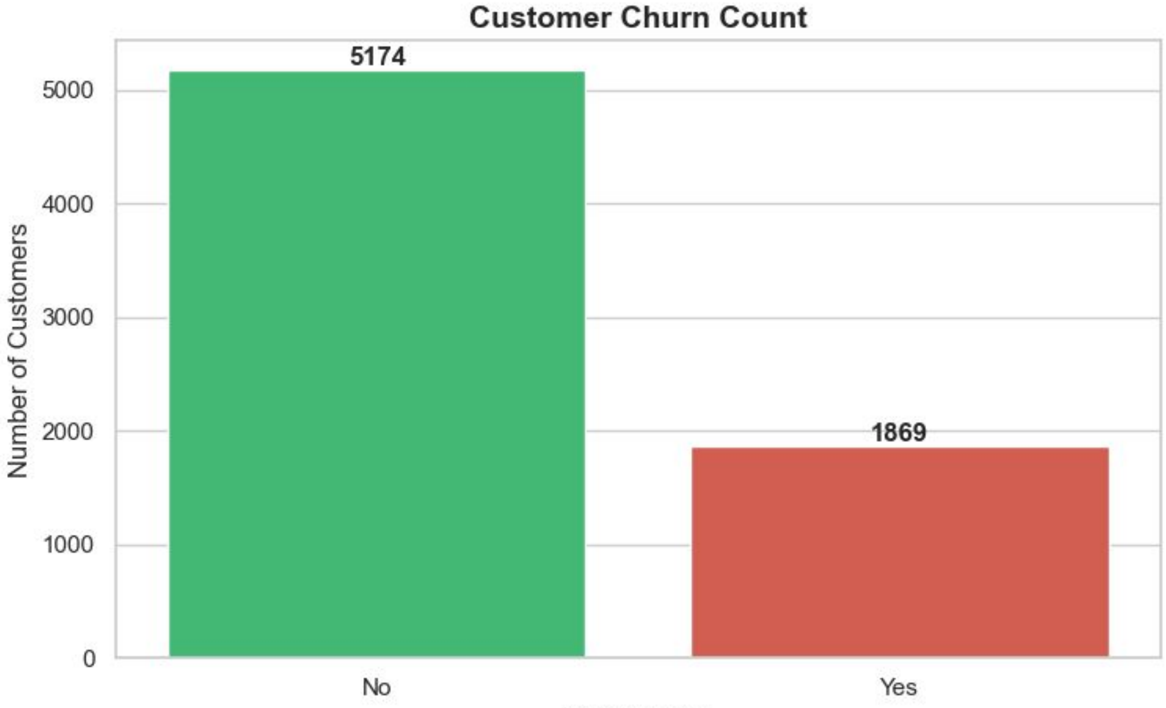
(Rows)

Source: kaggle.com/datasets/blastchar/telco-customer-churn/data

Exploratory Data Analysis – Target Distribution

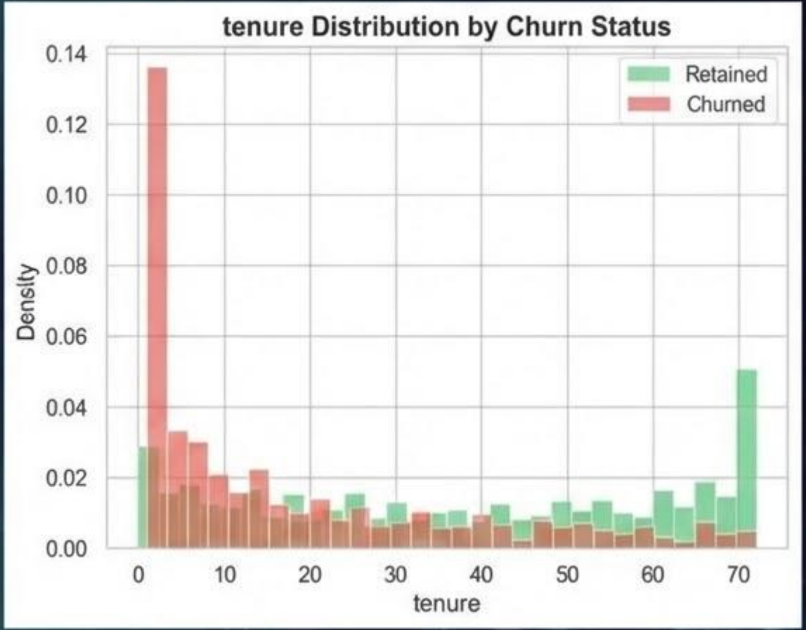
Dataset is imbalanced, with a 26.5% churn rate. This baseline informs the modeling strategy, requiring metrics like recall and ROC-AUC over simple accuracy.





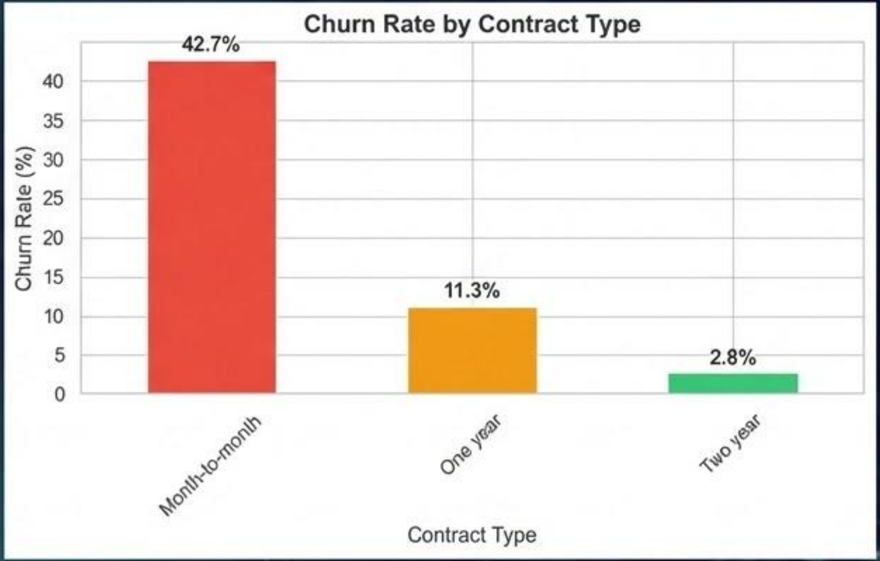
Key Churn Driver – Tenure

Churn risk is highest during the first 12 months

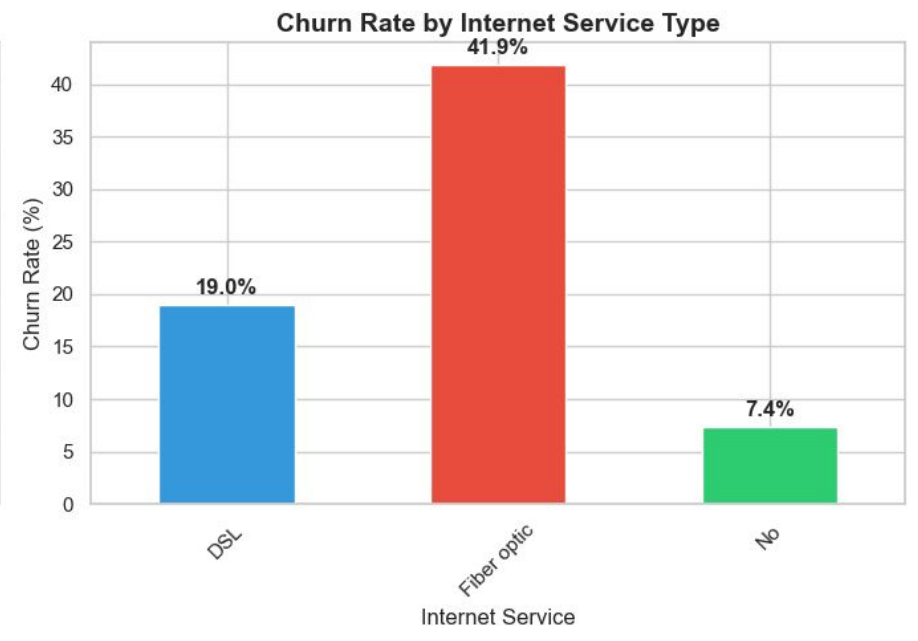


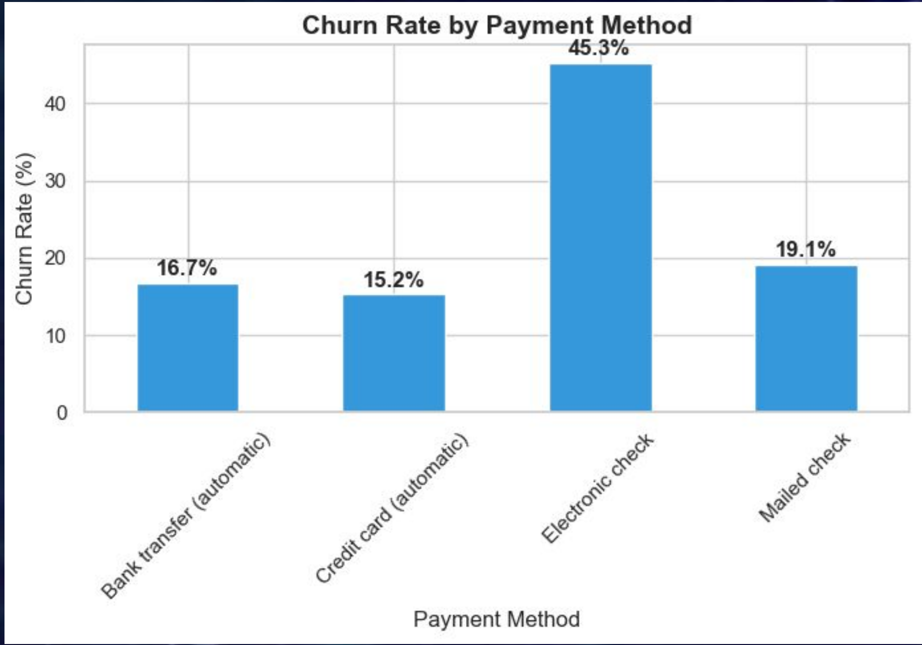
Key Churn Driver – Contract Type

Customers on month-to-month contracts are significantly more likely to leave compared to those on one- or two-year commitments



Additional Churn Drivers - Internet Service and Payment Method





Fiber optic service and electronic check payments are high-risk indicators.

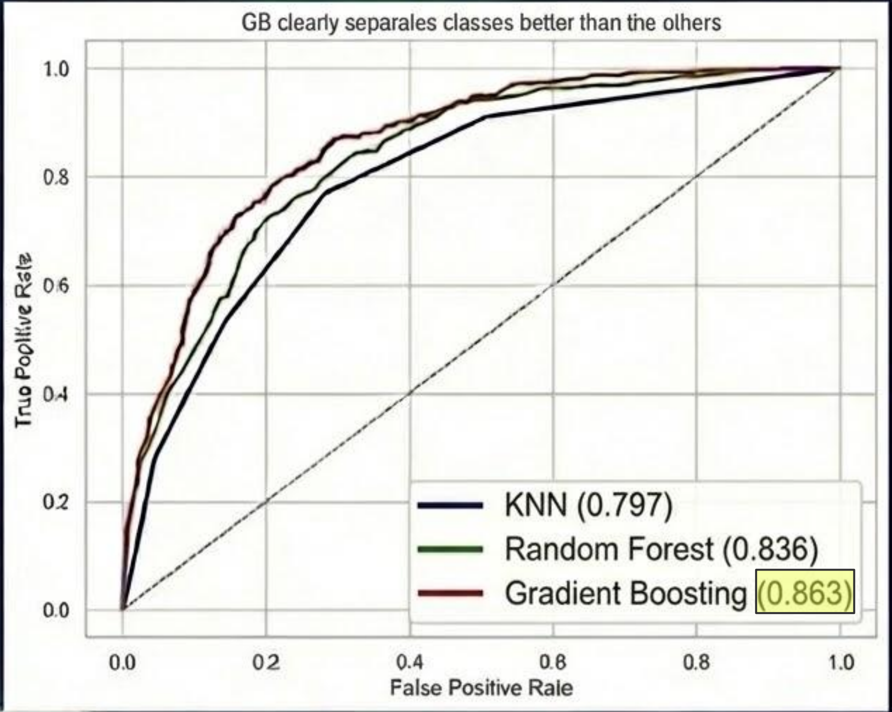
Data Preprocessing & Feature Engineering

Step	Action Taken	Details & Rationale
Missing Value Treatment	Imputed 11 records for `TotalCharges`	Filled empty values (new customers with 0 tenure) with `MonthlyCharges` to retain data integrity.
Feature Encoding	Dummy Encoding via `pd.get_dummies()`	Converted categorical variables into 30 features. Used `drop_first=True` to prevent multicollinearity.
Feature Scaling	Standardization via `StandardScaler`	Scaled `tenure`, `MonthlyCharges`, and `TotalCharges` to ensure equal weighting for distance-based algorithms.
Train/Test Split	80/20 Random Split	Divided data into Training (80%) and Testing (20%) sets using a fixed `random_state` for reproducibility.
Data Validation	Distribution Verification	Manually verified that the Churn rate (approx. 26.5%) remained consistent across all sets to ensure fair representation.

Modeling Strategy

Model	Rationale
K-Nearest Neighbors	Simple baseline; effective for pattern recognition.
Random Forest	Selected for its effectiveness with tabular, categorical data.
Gradient Boosting	Chosen for its superior predictive power through iterative error reduction.

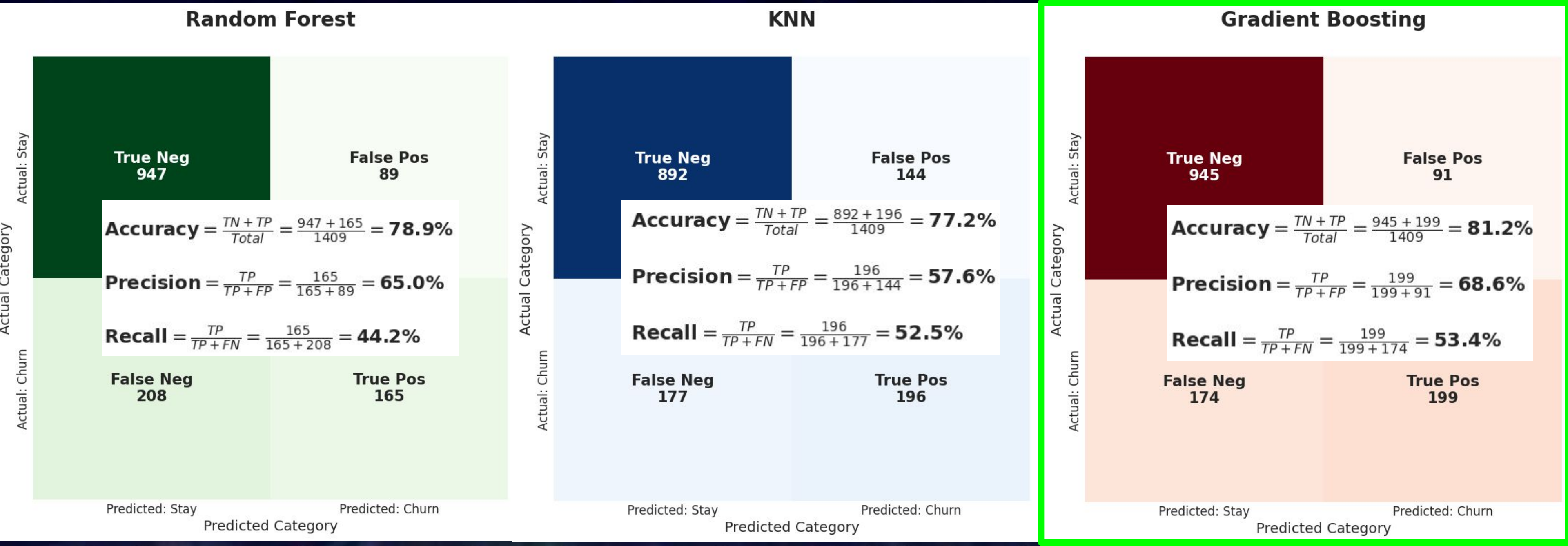
Baseline Performance (ROC Curve)



Model	CV Mean	CV Std
KNN	0.7829	0.0082
Random Forest	0.8227	0.0124
Gradient Boosting	0.8416	0.0106

Ran 5-fold CV to check that these results weren't just due to a lucky train/test split.

Model Performance (Confusion Matrix)



Gradient Boosting was selected because it provided the best balance of identifying the most at-risk customers while maintaining the lowest rate of false alarms.

Hyperparameter Tuning

Gradient Boosting showed the best baseline performance, but I tuned both ensemble models to see if I could close the gap or improve further.

Gradient Boosting

- ‘learning_rate’: 0.05
- ‘max_depth’: 3
- ‘n_estimators’: 100
- ‘subsample’: 0.8

Random Forest

- ‘max_depth’: 10
- ‘min_samples_leaf’: 2
- ‘min_samples_split’: 5
- ‘n_estimators’: 200

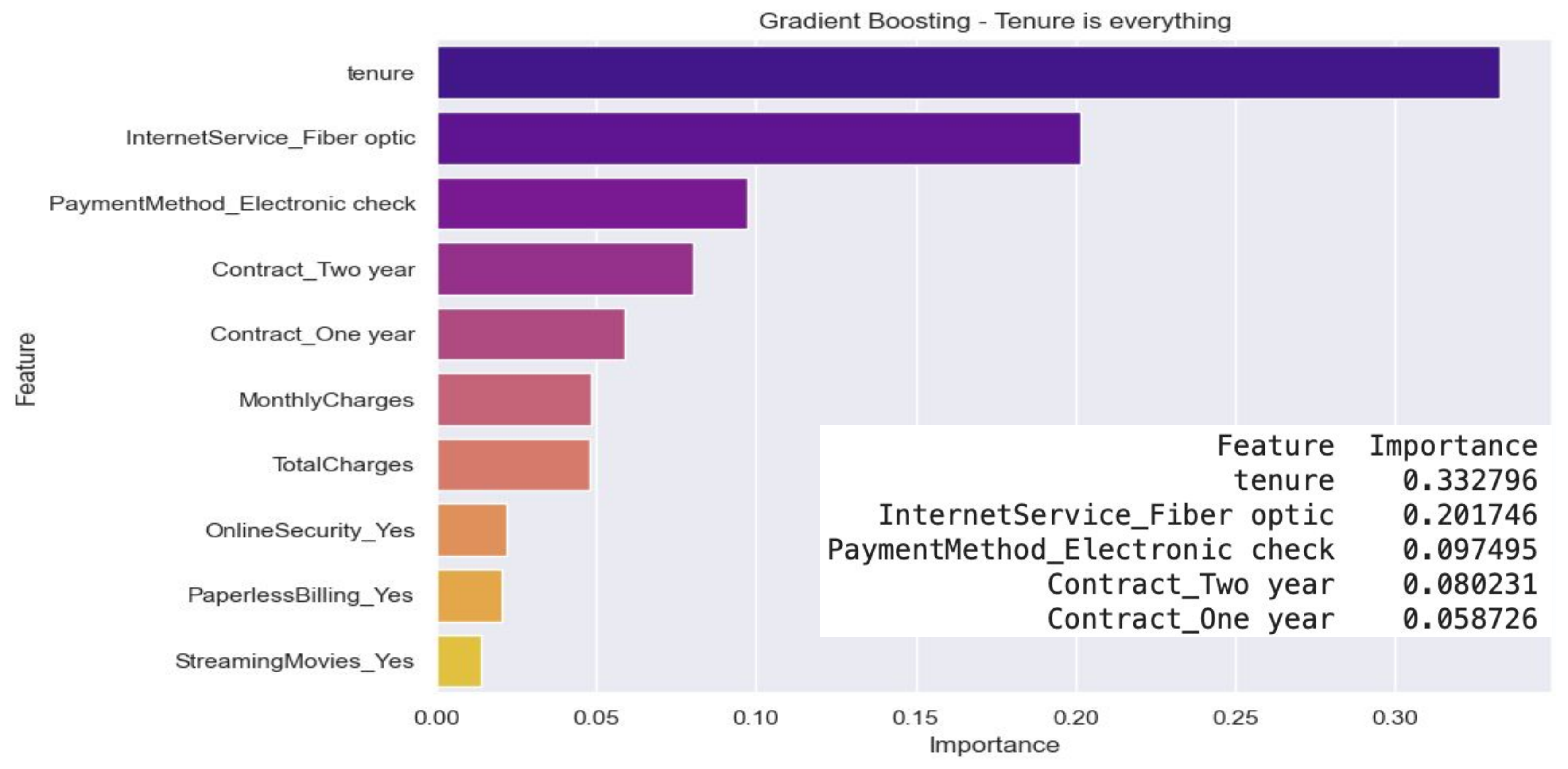
Used a systematic Grid Search with 5-fold cross-validation to identify this optimal hyperparameter set.

Tuned Model Comparisons

Gradient Boosting (Tuned) achieved the best ROC-AUC (0.865) and highest recall, making it my recommended model.

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
KNN	0.7722	0.5765	0.5255	0.5498	0.7970
Random Forest	0.7892	0.6496	0.4424	0.5263	0.8358
Random Forest (Tuned)	0.8169	0.7076	0.5255	0.6031	0.8632
Gradient Boosting	0.8119	0.6862	0.5335	0.6003	0.8630
Gradient Boosting (Tuned)	0.8148	0.6972	0.5308	0.6027	0.8654

Feature Importance



Feature	Importance
tenure	0.332796
InternetService_Fiber optic	0.201746
PaymentMethod_Electronic check	0.097495
Contract_Two year	0.080231
Contract_One year	0.058726

Conclusion

This project successfully developed a machine learning model to predict customer churn with **86.5%** ROC-AUC, exceeding the 80% target.

Tenure is paramount

The **first year** is **make-or-break** for customer retention

Contracts drive loyalty

Longer commitments correlate with dramatically **lower churn**

Service choices matter

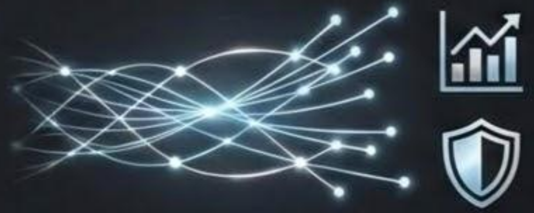
Fiber optic and **electronic check payment** are risk indicators

Business Recommendations

1. **Implement a “First 90 Days” Retention Program**
   - Check-in calls at 30, 60, and 90 days
   - Offer loyalty incentives at the 6-month and 12-month marks

2. **Incentivize Contract Commitments**
   - Offer discounts for one-year and two-year contracts
   - Create a “contract conversion" campaign targeting long-tenured month-to-month customers

3. **Investigate and Address Fiber Optic Service Issues**
   - Analyze service quality metrics and complaint data
   - Consider service level guarantees or credits for outages



Future Research

Survival Analysis

Optimize intervention timing.

Customer Lifetime Value Integration

Prioritize high-value retention efforts.

Natural Language Processing

Identify early warning signals.

A/B Testing Framework

Measure effectiveness of interventions.

Segment-Specific Models

Address different churn patterns.